# DETAILED PHENOTYPING AND NEXT-GENERATION SEQUENCING FOR CHARACTERIZATION OF RARE OVERGROWTH SYNDROMES

by

Ana Sequerra Amram Cohen

B.Sc., Genetics, University of Glasgow, 2010

M.Res., Biomedical Sciences, University of Glasgow, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Medical Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

January 2017

## Abstract

Weaver syndrome (WS) is a rare overgrowth disorder characterized by tall stature, macrocephaly, advanced bone age, facial dysmorphism, intellectual disability and cancer susceptibility, and it is caused by constitutional mutations in the enhancer of zeste homolog 2 gene (*EZH2*). To expand our understanding of WS pathogenesis, we assembled a cohort of 66 individuals with Weaver-like features, and collected DNA together with detailed clinical information. Sanger sequencing identified eleven individuals with pathogenic mutations in *EZH2* (equivalent to a 17% diagnostic rate). A further seven individuals carried mutations in the nuclear receptor-binding SET domain-containing protein 1 gene (*NSD1*), which cause a similar overgrowth disorder called Sotos syndrome (11% diagnostic rate). Furthermore, we expanded the phenotypic spectrum of WS to include neuronal migration disorders.

EZH2 is a histone methyltransferase that acts as the catalytic agent of the Polycomb Repressive Complex 2 (PRC2) to maintain gene repression via methylation of lysine 27 on histone H3 (H3K27). Functional studies investigating the activity of mutant EZH2 from various cancers showed that both gain- and loss-of-function mechanisms exist, thus it was important to determine which mechanism is causing WS. Using a standard histone methyltransferase assay, we observed that WS-associated *EZH2* mutations impair PRC2's histone methyltransferase activity *in vitro*, suggesting a loss-of-function mechanism of disease. In addition, no correlation between degree of functional impairment and phenotypic severity was noted.

Recognizing a clear role for chromatin modifications in the molecular pathophysiology of overgrowth syndromes, we hypothesized that mutations in other chromatin regulators might explain the phenotype observed in the remaining undiagnosed individuals. Using next-generation sequencing in combination with detailed phenotyping, we identified *EED* as a novel overgrowth gene. EED happens to be the main partner of EZH2 within PRC2, and is necessary for proper H3K27 methylation to occur.

Altogether, we have expanded the phenotypic and mutational spectrums of WS, and begun to uncover the underlying mechanism of disease. We also discovered a novel overgrowth gene, *EED*, reinforcing a role for PRC2 in the regulation of human growth and development.

## Preface

This thesis is comprised of both published and unpublished material as described below. This study was carried out under the supervision of Dr. William Gibson at the British Columbia Children's Hospital Research Institute (formerly Child & Family Research Institute), and it was approved by the joint Clinical Research Ethics Board of the University of British Columbia (UBC) and British Columbia Children's Hospital (BCCH); certificate numbers H08-00784, H09-01228 and H10-03215.

Chapter 1 is a literature review compiled for this thesis, and not published elsewhere.

Chapter 2 consists mainly of unpublished work characterizing the cohort of this study. However, the genotyping results and phenotypic tables for probands 3-7 and cases 15, 40, 53, 73, and 95, as well as part of this chapter's conclusions, were previously published in: Cohen ASA, Yap DB, Lewis MES, Chijiwa C, Ramos-Arroyo MA, Tkachenko N, Milano V, Fradin M, McKinnon ML, Townsend KN, Xu J, Van Allen MI, Ross CJD, Dobyns WB, Weaver DD, Gibson WT (2016) **Weaver syndrome-associated EZH2 protein variants show impaired histone methyltransferase function in vitro**. *Human Mutation* 3: 301-307. [PMID: 26694085].

I collected the patients' phenotypic information from referring physicians, coordinated the sample collection, carried out the DNA extractions, developed and conducted the sequencing experiments, analyzed and interpreted the Sanger sequencing results, wrote the manuscript, generated the figures and tables, and addressed the reviewers' comments with input from the other contributing authors.

Contribution from co-authors (for sections included in this Chapter only):

- Dr. Lewis, Chieko Chijiwa, Dr. Ramos-Arroyo, Dr. Tkachenko, Dr. Milano, Dr. Fradin, Dr. McKinnon, and Dr. Van Allen provided patient data;
- Dr. Weaver provided patient data and also advised on all of the patients' written descriptions;
- Dr. Dobyns reviewed the MRI scans for proband 5;
- Dr. Colin Ross provided guidance in setting up the sequencing protocols, as well as access to the Sanger sequencing core facility at the Centre for Molecular Medicine and Therapeutics;

- Katelin Townsend (research technician) and Jieqing Xu (summer student under my supervision) assisted me with the initial Sanger sequencing set-up and analysis;
- Dr. William Gibson provided in-depth edits for the manuscript and overall guidance.

Chapter 3 is based on my work conducted at the British Columbia Cancer Research Centre under the guidance of Dr. Damian Yap from the Aparício laboratory. A version of this work has been published in: Cohen ASA, Yap DB, Lewis MES, Chijiwa C, Ramos-Arroyo MA, Tkachenko N, Milano V, Fradin M, McKinnon ML, Townsend KN, Xu J, Van Allen MI, Ross CJD, Dobyns WB, Weaver DD, Gibson WT (2016) **Weaver syndrome-associated EZH2 protein variants show impaired histone methyltransferase function in vitro**. *Human Mutation* 3: 301-307. [PMID: 26694085].

I conducted all of the experiments and data analysis, wrote the manuscript, generated the figures and tables, and addressed the reviewers' comments with input from the other contributing authors, as well as from Dr. Gregg Morin (UBC). The methods presented in this chapter were adapted from Dr. Yap's original publication: Yap *et al*. (2011). Somatic mutations at EZH2 Y641 act dominantly through a mechanism of selectively altered PRC2 catalytic activity, to increase H3K27 trimethylation. *Blood* 117: 2451-2459.

Contribution from co-authors (for sections included in this Chapter only):
- Dr. Damian Yap provided direct mentorship in experimental set-up and data analysis;
- Dr. William Gibson provided in-depth edits for the manuscript and overall guidance.

Chapter 4 includes a combined version of two case reports that have been previously published (section 4.3). The first case report represents the first description of germline mutations in *EED* as a novel cause for overgrowth: Cohen ASA, Tuysuz B, Shen Y, Bhalla SK, Jones SJM, Gibson WT (2015). **A novel mutation in *EED* associated with overgrowth**. *Journal of Human Genetics* 60: 339-342. [PMID: 25787343]. Exome sequencing was conducted at the Canada's Michael Smith Genomes Sciences Centre (GSC). I carried out the sequencing validations, analyzed and interpreted the Sanger sequencing results, wrote the manuscript, generated the figures and tables, and addressed the reviewers' comments with input from the other contributing authors.

Contribution from co-authors (for sections included in this Chapter only):

- Dr. Tuysuz provided patient data;
- Yaoqing Shen executed the bioinformatic analysis of the raw exome data and generated a list of rare variants at the GSC, in which she noted the presence of a variant in *EED*;
- Dr. Bhalla reviewed the X-ray images of this patient;
- Dr. Steven Jones provided access to the GSC sequencing facilities and scientific collaboration for this ongoing study;
- Dr. William Gibson provided in-depth edits for the manuscript and overall guidance.

The other publication combined into section 4.3 of Chapter 4 reports on the second case of *EED*-related overgrowth: Cohen ASA and Gibson WT (2016). ***EED*-associated overgrowth in a second male patient**. *Journal of Human Genetics* 61: 831-834. [PMID: 27193220]. I developed and conducted all of the experiments, analyzed and interpreted the Sanger sequencing results, wrote the manuscript, generated the figures and tables, and addressed the reviewers' comments. My supervisor Dr. William Gibson carried out a detailed review of the patient's phenotype and provided in-depth edits for the manuscript and overall guidance.

Chapter 5 provides a discussion of my thesis work, and has not been published elsewhere.

## Table of Contents

## List of Tables

## List of Figures

## List of Abbreviations

| | |
|---|---|
| $^{3}H$ | Tritium (radioactive label) |
| ADD | *ATRX, DNMT3, DNMT3L* domain |
| AEBP2 | Adipocyte enhancer-binding protein |
| AML | Acute myeloid leukemia |
| AWS | *Associated with SET* domain |
| BCCH | British Columbia Children's Hospital |
| bp | Base pair(s) |
| BWS | Beckwith-Wiedemann syndrome |
| CBP | Gene name CREBBP: CREB-binding protein |
| CDC | Centers for Disease Control and Prevention |
| CHD3 | Chromodomain helicase DNA-binding protein 3 |
| ChIP-seq | Chromatin immunoprecipitation (ChIP) with next-generation sequencing |
| ClinVar | Clinical Variation (database) |
| CMMT | Centre for Molecular Medicine and Therapeutics |
| CNV | Copy number variant |
| COMPASS | Complex of Proteins Associated with Set1 |
| COSMIC | Catalogue of Somatic Mutations in Cancer |
| CXC | Cysteine rich motif |
| dbSNP | Database of short genetic variations |
| DECIPHER | Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources |
| DMSO | Dimethyl sulfoxide |
| DNA | Deoxyribonucleic acid |
| DNMT | DNA methyltransferase |
| DNMT3A | DNA methyltransferase 3A |
| DTT | Dithiothreitol |
| EDTA | Ethylene diamine tetraacetic acid |
| EED | Embryonic ectoderm development protein |
| ESC | Embryonic stem cell |

| | |
|---|---|
| ESP | Exome Sequencing Project |
| EVS | Exome Variant Server |
| ExAc | Exome Aggregation Consortium |
| ExAc | Exome Aggregation Consortium |
| EZH2 | Enhancer of zeste, homolog 2 |
| GSC | Canada's Michael Smith Genomes Sciences Centre |
| H2AK119 | Lysine residue 119 on histone H2A |
| H3K4 | Lysine residue 4 on histone H3 |
| H3K4me3 | Trimethylated H3K4 |
| H3K27 | Lysine residue 27 on histone H3 |
| H3K27ac | Acetylated H3K27 |
| H3K27me0 | Unmethylated ("naïve") H3K27 |
| H3K27me1 | Monomethylated H3K27 |
| H3K27me2 | Dimethylated H3K27 |
| H3K27me3 | Trimethylated H3K27 |
| H3K36 | Lysine residue 36 on histone H3 |
| H3K36me2 | Dimethylated H3K36 |
| H3K36me3 | Trimethylated H3K36 |
| HAT | Histone acetyltransferase |
| HDAC | Histone deacetylase |
| HDM | Histone demethylase |
| HMT | Histone methyltransferase |
| Hox/HOX | Homeotic (genes) |
| HPO | Human Phenotype Ontology (terms) |
| IMAGE | Intrauterine growth retardation, metaphyseal dysplasia, adrenal hypoplasia congenital, and genital anomalies |
| indel | Insertion/deletion (for small insertion or deletion) |
| IQ | Intelligence quotient |
| JARID2 | Jumonji and AT-rich interaction domain containing 2 |
| JMJD3 | Jumonji domain-containing protein 3 (also known as KDM6B: lysine-specific demethylase 6B) |

| | |
|---|---|
| kb | Kilo base pair(s) |
| LC | "Literature" cohort |
| LOVD | Leiden Open (source) Variation Database |
| MAF | Minor Allele Frequency |
| MEGF8 | Multiple epidermal growth factor-like domains 8 |
| MRI | Magnetic resonance imaging |
| NC | "*EZH2*/*NSD1* mutation negative" cohort |
| NCBI | National Center for Biotechnology Information |
| NHLBI | National Heart, Lung, and Blood Institute |
| NSD1 | Nuclear receptor-biding SET domain protein 1 |
| NuRD | Nucleosome remodeling and histone deacetylase |
| OMIM | Online Mendelian Inheritance in Man |
| p300 | Gene name EP300: E1A-binding protein, 300-KD |
| PcG | Polycomb group genes |
| PCR | Polymerase chain reaction |
| PHD | Plant homeodomain (zinc finger motif) |
| Polyphen | Polymorphism Phenotyping |
| PRC1 | Polycomb Repressive Complex 1 |
| PRC2 | Polycomb Repressive Complex 2 |
| PROVEAN | Protein Variation Effect Analyzer |
| PWWP | Pro-Trp-Trp-Pro domain |
| RBAP48 | Gene name RBBP4: retinoblastoma-binding protein 4 |
| RT-PCR | Reverse Transcription Polymerase Chain Reaction |
| S.D. | Standard Deviations |
| SAM | S-adenosyl-methionine |
| SANT | *Swi3, Ada2, N-Cor, and TFIIIB* domain |
| SET | *Su(var)3-9, E(z) and Trithorax* domain |
| SETD2 | SET-domain containing protein 2 |
| SGBS | Simpson-Golabi-Behmel syndrome |
| SIFT | Sorting Intolerant From Tolerant |
| SNP | Single nucleotide polymorphism |

| | |
|---|---|
| SNV | Single nucleotide variants |
| SS | Sotos syndrome |
| SUZ12 | Suppressor of zeste 12 (homolog) |
| Tris-HCl | Tris hydrochloride |
| TrxG | Trithorax group genes |
| UBC | University of British Columbia |
| UTR | Untranslated region |
| UTX | Ubiquitously transcribed tetratricopeptide repeat gene on X chromosome (also known as KDM6A: lysine-specific demethylase 6A) |
| VOUS | Variant(s) of unknown significance |
| WD | Tryptophan-aspartic acid dipeptide |
| WD40 | WD-domain (tandem copies of WD dipeptides) |
| WES | Whole exome sequencing |
| WGS | Whole genome sequencing |
| WS | Weaver syndrome |
| WT | Wild-type |
| Xi | Inactive X chromosome |
| XIST | X inactivation-specific transcript |

## Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. William Gibson, for the opportunity to carry out this engaging project in his lab, and for his continuous guidance and support throughout my PhD.

Next, I thank the members of my Supervisory Committee, Dr. Jan Friedman, Dr. Steven Jones, and Dr. Brad Hoffman, for their insightful comments and expert suggestions, and I thank Dr. Damian Yap for his mentorship in carrying out the functional studies presented in this thesis.

I would also like to recognize the other members of the Gibson lab, and fellow Medical Genetics students at UBC, for the constant encouragement, inspiration, and lively scientific discussions; and Cheryl Bishop, who provided constant administrative and moral support.

Further, this project would have not been possible without the patients who participated in our study, as well as their family members, physicians and genetic counsellors that contributed valuable information.

An immense gratitude goes to the Fundação para a Ciência e a Tecnologia (FCT) in Portugal, who provided me with a Doctoral Grant to carry out the research presented here.

Finally, a special thanks to my family and friends who have supported me throughout my years of education away from my home country, with particular mention of my mother Lia, who has given me unconditional encouragement and all the emotional support I needed to complete this doctoral thesis.

This thesis is dedicated to two of my biggest supporters, to whom I had to say goodbye during my PhD studies. They were exceptional beings with hearts of gold, and gave me unlimited love and encouragement throughout my young life. They showed me how important it is to care for others and inspired me to work hard and follow my dreams.

*You will always be remembered, Salomão & António.*

# Chapter 1: Introduction

## 1.1 Rare overgrowth syndromes

A rare disease is defined as one that affects fewer than 200,000 persons in the United States, or fewer than 1 in 2,000 people in Europe.[1] Although individually rare, the estimated 7,000 rare monogenic diseases collectively affect 1 in 50 individuals.[2,3]

Overgrowth syndromes are a subset of rare genetic disorders characterized by excessive growth, which may be localized or diffuse. In recent years, localized overgrowth was confirmed to be caused by somatic mosaicism,[4–6] whereas generalized overgrowth is thought to be caused by constitutional alterations of the genome or epigenome.[7] Chromosomal deletions and duplications that cause overgrowth are usually picked up by clinical microarray, while other alterations require further investigations.[8]

These rare overgrowth syndromes have many overlapping features, and typically manifest with tall stature, macrocephaly, and/or obesity, prenatally or shortly after birth, and in combination with other congenital features.[7,9] The most frequently observed overgrowth syndromes are Beckwith-Wiedemann syndrome (BWS) (OMIM #130650) and Sotos syndrome (SS) (OMIM #117550). The other best-defined overgrowth syndromes are Simpson-Golabi-Behmel syndrome (SGBS) (OMIM #312870) and Weaver syndrome (WS) (OMIM #227590). The latter represents the focus of this thesis.

### 1.1.1 Weaver syndrome

In 1974, Weaver *et al*.[10] described two "strikingly similar" unrelated young boys with generalized overgrowth of prenatal onset, accelerated osseous maturation, hypertonia, camptodactyly, hoarse low-pitched cry, and a specific combination of craniofacial features (including large ears, broad forehead with hypertelorism, and long philtrum with relative micrognathia). Other overlapping features were also noted, such as: excessive appetite, limited elbow and knee extension, widened distal femurs and ulnas, prominent finger pads and broad thumbs with thin deep-set nails, clinodactyly in the toes, excessive loose skin, hypoplastic nipples, thin hair, umbilical and inguinal hernias. The combination of these phenotypic traits was

defined as a novel overgrowth disorder named Weaver syndrome (WS), and these two cases represent the "classical" Weaver presentation against which future cases were compared.

Following the initial description of WS, over 70 clinical cases have been reported in the literature.[11–54] In analysing the clinical presentation retrospectively, at least 63 of these cases are consistent with a clinical diagnosis of WS, with slightly more males affected than females (approximately 59% males to 41% females). Seeing that both sexes are affected, an X-linked genetic cause was excluded. Notably, most published reports were of affected children with no family history, suggesting that WS was caused by new dominant mutations.

In 1992, Cole *et al*.[31] noted that classical Weaver features are more prominent in early childhood and become less prominent in adulthood; a milder phenotype could partially explain the lower number of adult cases reported. As such, we recommend that photographs from early childhood be used to aid in the diagnosis of adolescents and adults. This milder phenotype in adulthood could also be masking a clinical diagnosis in some parents of the apparently sporadic cases reported. Nonetheless, there were a few familial cases reported,[19,24,34,36,38,41,43,51] and both paternal and maternal inheritance were observed, suggesting that an imprinting effect was unlikely. These familial cases supported an autosomal dominant mode of inheritance.

As the number of reports increased, some phenotypic variability became apparent. Generalized overgrowth remained a hallmark, but it was not always of prenatal onset and was variable in degree. Similarly, the "classical" Weaver combination of facial features remained necessary for diagnosis but not all features were as prominent in every individual. Further, the advanced bone age, camptodactyly and hoarse low-pitched cry that had been described in the original report [10] were not observed in all cases, and neither was hypertonia; some individuals presented with hypotonia instead, suggesting that abnormal muscle tone may remain a diagnostic criteria. In addition, the following features were described in few (or even single) cases: pes cavus,[18] instability of the upper cervical spine,[26] respiratory disorders,[27] cardiac abnormalities,[39] neuronal migration disorders,[40,55] dental problems,[48,54] and congenital hypothyroidism.[49] Moreover, it was rapidly appreciated that patients with WS also have a predisposition to malignancies, with an incidence of 6/63 (9.5%) amongst the clinical cases described.[26,41,43,44,50,52]

The underlying genetic cause for this classical overgrowth disorder had remained unsolved until the age of exome sequencing. It wasn't until late 2011 that Gibson *et al*. and others showed that mutations in the enhancer of zeste homolog 2, or *EZH2* (OMIM *601573), cause WS.[56,57] To

the best of our knowledge, 57 cases of molecularly confirmed WS have since been published including seven of our own,[55–60] with 25 males (44%) and 32 females (56%) reported. Of note, two of the variants listed as causative by Tatton-Brown *et al*.[58] appear to be intronic variants of unknown inheritance, which we would consider "probably pathogenic". Of these 57 cases, six developed malignancies (10.5%);[44,58–60] this high rate is likely reflective of ascertainment bias, as clearly these individuals are more severely affected and thus more likely to be reported in the literature. Also, we do not know whether any of these cases overlap with previously published cases of clinically diagnosed WS (except for our own probands 1 and 4, as described later), and therefore we cannot accurately estimate the number of cases reported. WS is generally caused by *de novo* mutations, thus explaining the high number of isolated sporadic cases reported, although over 25% of molecularly confirmed cases reported by Tatton-Brown *et al*. in 2013 are familial.[58] Pathogenic variants are distributed throughout the coding region of *EZH2* and not exclusively within recognizable protein domains,[55–60] as summarized in Figure 1-1. The vast majority of mutations are missense, with some being recurrent.[55–61]


**Figure 1-1: Distribution of Weaver syndrome mutations across *EZH2*.**

Human EZH2 is represented. Each rectangle represents one exon. Exon size is represented to scale, whereas intronic distances are not to scale. White (open) rectangles represent non-coding UTRs and grey rectangles represent coding exons (NM_004456.4). EZH2 protein contains 751 amino acids (NP_004447.2) and five recognizable domains (two SANT, and one each of WD-binding, CXC and SET), represented here in coloured rectangles according to NCBI (NP_004447.2) and UniProtKB/InterProt (Q15910-2) coordinates. Each triangle represents one mutation identified in a Weaver syndrome patient as reported by Gibson *et al*. and Cohen *et al*. (in red),[56,59] Tatton-Brown *et al*. (in blue/orange),[58] Al-Salem *et al*. (in yellow),[55] and Usemann *et al* (in green).[60] Unpublished mutations from our cohort are also represented in red. e = exon; UTR = untranslated region; SANT = *Swi3, Ada2, N-Cor, and TFIIIB* domain; CXC = Cysteine rich motif; SET = *Su(var)3-9, E(z) and Trithorax* domain.

### 1.1.2 Sotos syndrome

Weaver syndrome (WS) shares many features with Sotos syndrome (SS), and the two disorders are similar enough that even some dysmorphology specialists have difficulty distinguishing between the two.[61] Table 1-1 shows a comparison of the main characteristics between the two syndromes, based on scientific and medical literature.[61,62]

| | **Weaver syndrome** | **Sotos syndrome** |
|---|---|---|
| **General** | Overgrowth<br>Tall stature<br>Accelerated osseous maturation<br>Intellectual disability (milder) | Overgrowth<br>Tall stature<br>Accelerated osseous maturation<br>Intellectual disability |
| **Head** | Macrocephaly (~50%)<br>Large bifrontal diameter<br>Flat occiput | Macrocephaly (<50%)<br>Prominent forehead<br>Fronto-temporal hair sparsity |
| **Face** | Round (early years) | Long and thin |
| **Eyes** | Hypertelorism | Apparent hypertelorism<br>Down-slanting palpebral fissures |
| **Ears** | Large and fleshy | Large |
| **Chin** | Prominent chin crease ("stuck-on" chin)<br>Receding jaw (micro/retrognathia) | Prominent chin |
| **Others** | Cerebral malformations<br>Cancer predisposition (5-10%)<br>Umbilical hernias<br>Excessive "doughy" skin<br>Hoarse low-pitch cry<br>Long philtrum<br>Wide nasal root<br>Congenital cardiac anomalies (less frequent) | Large hands and feet<br>Cancer predisposition (~3%)<br>Behavioural issues (mainly autism)<br>Seizures<br>Congenital cardiac anomalies |

**Table 1-1: Clinical presentation of Weaver syndrome vs. Sotos syndrome.**

In 1964, ten years before the first report of WS, Sotos *et al*.[63] described five unrelated children that shared similar features of "cerebral gigantism" characterized by: excessively rapid growth including accelerated osseous maturation, non-progressive cerebral disorder with intellectual disability, and facial features such as a high broad forehead and a prominent jaw. In 1967, Hook and Reynolds presented six more cases including concordant monozygotic twins;[64] they noted that all individuals had macrocrania (increased size of the skull), large hands and feet, low IQ, ventricular problems, and general clumsiness, as well as the features described previously.[63] Since then, over one thousand cases have been reported in the literature. Naturally, these reports have expanded the clinical spectrum to include ocular abnormalities,[65–68] cardiac

defects,[69–72] endocrine disorders,[73–81] spinal instability or scoliosis,[82–85] behavioural issues or seizures,[86–89] dental complications,[90–93] severe connective tissue laxity,[94] and more. It was noted that many of these features become less apparent with age,[95,96] just like in WS; yet the oldest case reported with classical Sotos features is a 63 year old woman.[97] Association of SS with malignancies was also observed;[97–114] cancer prevalence is estimated to be around 3%. The majority of cases described are sporadic, with few reports of familial cases,[64,75,78,115–127] which suggested an autosomal dominant mode of inheritance.

In 2002, mutations and microdeletions in the nuclear receptor-binding SET domain containing protein 1, or *NSD1* (OMIM *606681), were found to cause SS. Because microdeletions were observed, the mechanism of disease was predicted to be haploinsufficiency.[128] Microdeletions are more common in the Japanese population,[129] while intragenic loss-of-function mutations, primarily truncating, are more common amongst the non-Japanese.[46,130–133] Overall, 80-90% of individuals clinically diagnosed with SS have alterations in *NSD1*.[61,132,134] Further, a "Sotos-like" syndrome (SOTOS2; OMIM #614753) has been reported by Malan *et al*.[135] and is caused by mutations in *NFIX*.[136]

### 1.1.3 Other overgrowth syndromes

In the 1960s, Beckwith and Wiedemann independently reported a distinct overgrowth syndrome, currently known as Beckwith-Wiedemann syndrome (BWS).[137,138] The clinical presentation of BWS is heterogeneous, but usually characterized by macrosomia, macroglossia (which can lead to difficulties in breathing, feeding and/or speech), omphalocele or umbilical hernias, ear creases/pits, visceromegaly of intra-abdominal organs, and other congenital anomalies and dysmorphisms.[62,139] High rates of prematurity and infant mortality are still observed.[139] A predisposition to embryonal malignancies, and particularly Wilms tumours, is estimated at around 7.5% in the first 8-10 years of life,[62,139,140] and this frequency appears to be highly dependent on the molecular cause of disease;[141,142] tumours at later ages are rare.[139] Similarly to WS and SS, BWS is usually sporadic, but maternal transmission is observed in 10-15% of cases.[139] BWS is caused by an imbalance in gene dosage within an imprinted region at 11p15.5;[139,140] for each gene within this region, only one of the alleles is expressed, and this expression is dependent on the parent from which the allele was inherited. The imprinting mechanism in this region is complex and involves two different domains (IC1 and IC2); the main

genes involved are *IGF2* and *KCNQ1OT1*, which are expressed off the paternal allele, and *H19*, *KCNQ1*, and *CDKN1C*, which are expressed off the maternal allele.[139,140] Both genetic and epigenetic alterations have been found to disrupt the gene dosage in this region,[139,140] including loss-of-function mutations in *CDKN1C*.[143] Interestingly, similar alterations in this region can also lead to growth retardation, a somewhat "mirror" phenotype: methylation dysregulations can cause Silver-Russell syndrome (OMIM #180860),[144] and gain-of-function *CDKN1C* mutations can cause IMAGE syndrome (OMIM #614732).[145]

Simpson-Golabi-Behmel syndrome (SGBS), observed less frequently, is an X-linked recessive disorder caused by mutations or deletions in *GPC3*.[146] It was named in 1988 by Neri *et al*.[147] who recognized the phenotypic overlap between patients previously described in three independent reports.[148–150] Some females are mildly affected due to random X-inactivation.[151] SGBS is characterized by pre- and postnatal overgrowth, macrocephaly, coarse facial features, and other congenital anomalies;[62,147,151] an increased risk for embryonal tumours has been noted.[151] There is also a more severe form of this disorder, named SGBS type 2 (OMIM #300209), which is caused by mutations in *CXORF5/ORD1* and leads to early death.[151,152] More recently, mutations in *PIGA* have also been suggested to cause SGBS type 2.[153]

Other disorders associated with overgrowth which may be considered for differential diagnosis include:

- Fragile X syndrome (OMIM #300624), an X-linked dominant disorder associated with macrocephaly in early years, which represents one of the most common causes of intellectual disability and is caused by mutations in *FMR1*;[154,155]
- Marfan syndrome (OMIM #154700), which is associated with disproportionate tall stature, long limbs, dysmorphism and other congenital abnormalities, and is generally caused by mutations in *FBN1*;[156,157]
- Beals syndrome (OMIM #121050), a disorder similar to Marfan syndrome generally caused by mutations in *FBN2*;[158]
- Perlman syndrome (OMIM #267000), associated with macrosomia and caused by homozygous or compound heterozygous mutations in the *DIS3L2* gene.[159]

Furthermore, novel syndromes associated with overgrowth were recently identified through next-generation sequencing technologies and classified as separate entities based on the

molecular cause rather than just the clinical presentation. These recent gene discoveries are summarized in Table 1-2 below.

| Syndrome | OMIM reference | Gene involved | Main features |
|---|---|---|---|
| Tatton-Brown-Rahman | #615879 | *DNMT3A* | tall stature, facial dysmorphism and intellectual disability [160] |
| Luscan-Lumish | #616831 | *SETD2* | macrocephaly, intellectual disability, behavioural difficulties; generalized overgrowth with obesity and advanced bone age described by Luscan *et al*.,[161] but short stature described by Lumish *et al*.[162] |
| Tenorio | #616260 | *RNF125* | overgrowth, macrocephaly, and intellectual disability [163] |
| Kosaki Overgrowth | #616592 | *PDGFRB* | tall stature, prominent forehead, downslanting palpebral fissures (both patients described are females of Japanese descent)[164] |
| MDFPMR: macrocephaly, dysmorphic facies, and psychomotor retardation | #617011 | *HERC1* | persistent macrocephaly, dysmorphism, tall stature, global developmental delay (autosomal recessive) [165] |
| *not yet established* | - | *PPP2R5B, PPP2R5C, PPP2R5D* | tall stature, macrocephaly and intellectual disability [166] (note: mutations in *PPP2R5D* also associated with mental retardation without overgrowth – see OMIM #616355) |
| Thauvin-Robinet-Faivre (TROFAS) | #617107 | *FIBP* | overgrowth, intellectual disabilities and multiple congenital anomalies (autosomal recessive) [167,168] |
| *not yet established* | - | *TCF20* | mild intellectual disability, postnatal tall stature and macrocephaly, obesity and muscular hypotonia [169] |

**Table 1-2: Novel overgrowth syndromes described in recent literature.**

## 1.2 Chromatin regulators in overgrowth and cancer

Regulation of gene activity often occurs without alterations to the DNA sequence itself, but rather by influencing when and where genes are expressed.[170,171] To fit within the cell nucleus, DNA requires a three-dimensional scaffolding system called "chromatin", wherein DNA is physically wrapped around nucleosomes (the basic units of chromatin).[170,172] Loosely-wrapped DNA is expressed, while tightly-wrapped DNA becomes inaccessible to the transcriptional machinery, thereby closing down local gene expression.[170] This system is reversible, and it is controlled by a myriad of proteins termed chromatin regulators. Each nucleosome contains eight core histone proteins, typically two copies each of H2A, H2B, H3 and H4; the histone tails that stick out of this core are accessible for modification.[170,172] Post-translational modification of histones, which is carried out by a subset of chromatin regulators, represents one major mechanism of regulating gene expression.[170,172]

Many overgrowth syndromes are caused by disruptions in chromatin regulators.[7] Interestingly, these same chromatin regulators also play an important role in cancer,[173] which could explain why a higher predisposition to malignancies is observed in patients with overgrowth.[7] As mentioned earlier, Weaver and Sotos syndromes are caused by mutations in *EZH2* and *NSD1* respectively, which encode two such chromatin regulators that can modify the tails of core histones. They are both histone methyltransferases (meaning they can add methyl groups to residues on histone tails) with opposing effects (repressing vs. activating), yet are responsible for very similar disease phenotypes when mis-regulated.[61,174]

### 1.2.1 EZH2

#### 1.2.1.1 EZH2 and the Polycomb Repressive Complex 2

##### 1.2.1.1.1 Polycomb and Trithorax group genes

*EZH2* was first described in flies as *E(z)*, and categorized as a Polycomb group gene.[175] In *Drosophila melanogaster*, Polycomb group (PcG) and Trithorax group (TrxG) genes regulate the expression of homeotic (Hox) genes, which in turn determine appropriate anterior/posterior body patterning boundaries and thus regulate fly embryonic development.[176,177] Mutations affecting these genes lead to drastic transformations of body segments.[176,177] Genes belonging to these two groups work in opposing yet balanced systems that respond to spatial and temporal cues throughout development: TrxG genes maintain active expression of homeotic genes within their appropriate boundary domains, while PcG genes repress their expression outside of these domains.[176,177] Much of this regulation is achieved by influencing chromatin structure, thus affecting the expression of other genes in addition to homeotic genes.[176,178] Given the notable conservation of the homeotic gene system across species, it was postulated that a similar chromatin regulatory system would also exist in more complex organisms.[176] Furthermore, it was proposed very early on that this chromatin regulation was likely to play a role in differentiation and lineage maintenance of eukaryotic cells.[179]

Most PcG and TrxG members act within multi-protein complexes.[178] The two main PcG complexes described are the Polycomb Repressive Complex 1 and 2 (PRC1 and PRC2); TrxG members typically associate into COMPASS-like complexes.[180–182] In *Drosophila melanogaster*, PRC1 is composed of four core proteins: Polycomb (Pc), Polyhomeotic (Ph), Posterior Sex combs (Psc) and Sex combs extra (RING).[183] In mammals, there are numerous possible

assemblies of PRC1, due in part to the existence of multiple paralogues for each of these four core subunits.[183] Typically, human PRC1 contains one chromobox protein (CBX2/4/6/7/8, Pc homologues), one polyhomeotic-like protein (PHC1/2/3, Ph homologues), either BMI-1 or MEL18 (Psc homologues), and one E3 ubiquitin ligase (RING1A/1B).[183,184] These subunits play different roles in mediating gene repression, together with PRC2 proteins, as discussed in the following two sections.

#### 1.2.1.1.2 Regulation of gene expression via methylation of H3K27 within PRC2

*E(z)* shows strong sequence conservation all the way up to mammals, including human *EZH2*;[185,186] the most highly conserved region of *E(z)*/*EZH2* encodes a *Su(var)3-9, E(z) and Trithorax* (SET) domain.[186,187] To mediate gene repression, EZH2 acts as a histone methyltransferase (HMT) via the SET domain, with specificity for the histone H3 tail and strong preference for lysine residue 27 (H3K27).[178,188,189] EZH2 can add up to three methyl groups onto H3K27, producing mono- di- and trimethylated lysine (H3K27me1, H3K27me2 and H3K27me3, respectively) in a sequential manner (Figure 1-2).[190,191] H3K27me1 marks are usually found along the bodies of active genes, H3K27me2 marks locate to intergenic regions, and H3K27me3 marks are associated with repressed chromatin and found at promoters of silenced genes (where promoters are defined as a 5 kb region centered on the transcriptional start site).[192]

In order to mediate its HMT activity, EZH2 must be incorporated into PRC2, together with two other essential proteins: EED (embryonic ectoderm development protein; OMIM *605984) and SUZ12 (suppressor of zeste 12; OMIM *606245).[178,188,189] Other components of PRC2 can vary dependent on cellular context,[178,193] but RBAP48 is usually observed in the human complex (Figure 1-2).[188,189] The EZH2-EED partnership had been observed across species, being conserved from plants to humans, thus supporting a need for this interaction for proper HMT activity.[188,194] Later on, SUZ12 was shown to be likewise required for PRC2-mediated HMT activity to occur,[189] while RBAP48 and cofactors such as AEBP2 and JARID2 contribute to stimulate the enzymatic activity.[189,195–197] In some contexts, PRC2-mediated H3K27 methylation is reversible by H3K27-specific demethylases (HDMs) such as UTX and JMJD3 (see Figure 1-3).[181,198–200]

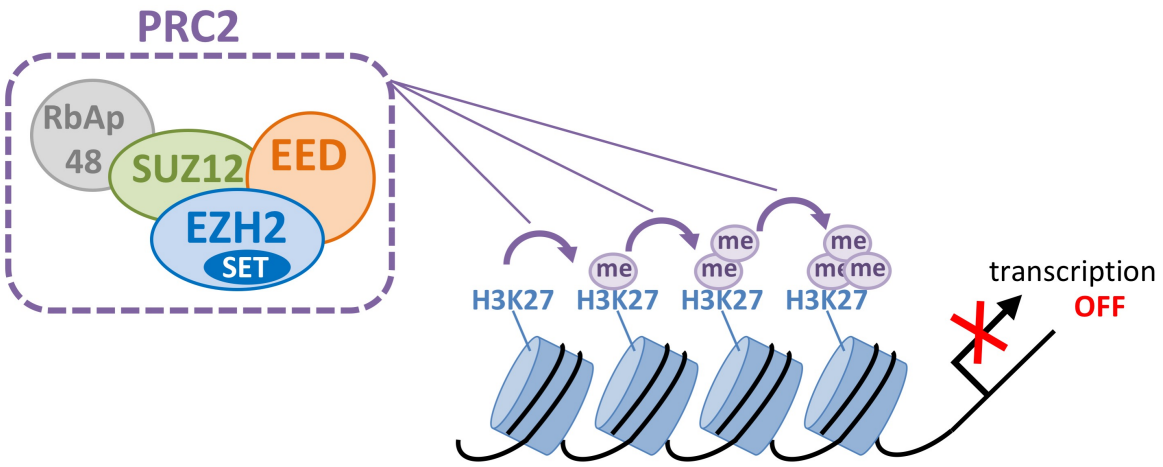



**Figure 1-2: Schematic representation of the histone methyltransferase function of EZH2 within PRC2.** Schematic representation of the Polycomb Repressive Complex 2 (PRC2). EZH2 is the catalytic member of this chromatin regulator complex. Along with EZH2, EED and SUZ12 are also required for proper histone methyltransferase (HMT) activity mediated by the SET domain of EZH2, while other members of the complex vary depending on cellular context. Within PRC2, EZH2 can add up to 3 methyl groups (me) to lysine 27 on the histone 3 tail (H3K27). This is done in a sequential manner (H3K27me0 to H3K27me1, H3K27me1 to H3K27me2, and H3K27me2 to H3K27me3), and the H3K27me3 mark is associated with gene repression. Disruption of EZH2 is thought to disturb this PRC2-mediated HMT activity, and thus to influence the expression of hundreds of downstream targets.

PRC2 activity is directly related to its ability and/or affinity to bind to the H3 tail. As mentioned above, binding affinity changes with different PRC2 conformations, and may be dependent on the specific EED isoform included in the complex.[201] Binding specificity is also influenced by nucleosome numbers [202] and the presence of linker histone H1.[202] Successful trimethylation of H3K27 promotes further PRC2 binding to propagate this repressive mark via EED recognition.[203–205] H3K27me3 can also stimulate recruitment of PRC1 to chromatin (with H3K27me3 recognition by CBX proteins), which in turn monoubiquitylates H2AK119 (lysine residue 119 on histone H2A) via its E3 ubiquitin ligase subunit to support local gene repression.[178,183,188,206] Moreover, EZH2 has been shown to interact with DNA methyltransferases (DNMTs) to further sustain a repressive chromatin structure.[207]

Conversely, PRC2 activity is inhibited when EZH2 is phosphorylated by AKT, CDK1/2, or p38 alpha,[208,209] when H3K27 is mutated genetically (for example by the H3.3K27M recurrent mutation observed in pediatric gliomas)[210–212] or modified biochemically (with the addition of an acetyl group),[213] and by the presence of other chromatin marks on the same H3 tail. Acetylation of H3K27 (H3K27ac) is carried out by histone acetyltransferases (HATs) such as CBP/p300, which compete with PRC2 to modify H3K27;[180,213] once established, an H3K27ac mark blocks the methylation of that residue (thus blocking gene repression), and so naturally it is associated with open chromatin and gene activation (see Figure 1-3).[180,213] This model is supported by the finding that knock-outs of essential PRC2 components in mouse embryonic stem cells (ESCs) led to a global loss of H3K27me3 and increase in H3K27ac levels, and re-expression of the missing protein restored H3K27me3 levels and decreased H3K27ac.[213] Histone deacetylases (HDACs) also provide an additional layer of antagonistic regulation: they reverse histone acetylation, thereby liberating the lysine residue for potential methylation.[204] In addition, there are other "activating" chromatin marks that can inhibit H3K27 methylation, including H3K4me3 (trimethylated lysine 4 on H3) put in place by SET and MLL proteins within COMPASS-like complexes,[182] and H3K36me2/me3 (di- and trimethylated lysine 36 on H3) added by TrxG protein ASH1 (see Figure 1-3).[214] Furthermore, some H3K27-specific HDMs have been shown to associate with HATs and H3K4-specific HMTs within COMPASS-like complexes to counteract PRC2's activity.[182,198] Overall, it appears that antagonizing PRC2 function is one of the main mechanisms through which TrxG proteins regulate gene expression.[182]

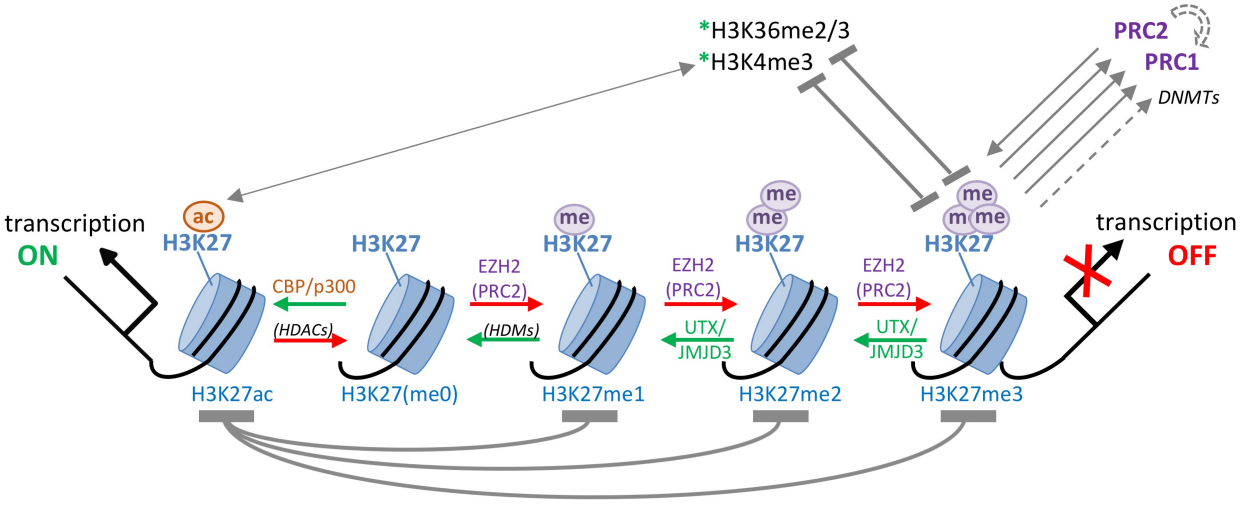



**Figure 1-3: Balance of Polycomb and Trithorax regulators in gene expression.**
Within the Polycomb complex PRC2, EZH2 is a histone methyltransferase (HMT) that adds up to three methyl groups (me) to H3K27. The H3K27me3 mark blocks transcription, and promotes the recruitment of further repressive agents (more PRC2, PRC1, and DNA methyltransferases or DNMTs). This methylation may be reversed by Trithorax-group demethylases (HDMs) such as UTX and JMJD3. PRC2's repressive activity is also repressed by other chromatin marks deposited on the H3 tail, often by Trithorax proteins, including acetylation of H3K27 (H3K27ac), and methylation of other lysine residues: H3K4me3 and H3K36me2/me3. In general, histone modifications mediated by Polycomb proteins (repressive) are represented in red, and modifications mediated by Trithorax proteins (activating) are represented in green. HDACs = histone deacetylases.

#### 1.2.1.1.3 Other biological functions of PRC2-mediated H3K27 methylation

As previously postulated, PRC2 and H3K27 methylation are implicated in various biological processes,[180,215] although the full extent of their involvement is not fully understood. The main role of PRC2 is in gene repression, which goes beyond homeotic genes as alluded to earlier. In fact, PRC2-mediated H3K27 methylation regulates the expression of hundreds of genes, and particularly genes related to cellular differentiation and transcription factors involved in development.[216,217] PRC2 targets across the genome were shown to be upregulated during mouse ESC differentiation,[216] suggesting that these genes are paused in the stem (pluripotent) state, waiting to become de-repressed during lineage commitment.[199,218] Such a "bivalent" state is dependent on simultaneous occupancy of nucleosomes with both activating (H3K4me3) and repressive (H3K27me3) marks, mediated by TrxG and PcG complexes respectively, which

allows genes to be turned on or off quickly as needed throughout differentiation (with HDMs removing either mark).[199,219] These opposing marks are thought to co-exist on the same nucleosome, but with one modification on each H3 tail, thus creating asymmetrically modified nucleosomes.[199,204,219] Fittingly, the pluripotency genes *Oct4* and *Nanog* were highly expressed in mouse ESCs, in correlation with low levels of H3K27me3, and became silenced upon cell differentiation.[216] These observations corroborate the involvement of PRC2 in determining cell fate by controlling the spatial and temporal expression of numerous genes necessary for establishing different cell and tissue types during development.[218,220] Further supporting the need for PRC2 in mammalian development is the finding that mouse knock-outs of *Ezh2*, *Eed* and *Suz12* are all embryonic lethal at early stages.[221–223] Additionally, ESCs lacking any of the three core PRC2 components failed to differentiate in culture,[220] and all states of methylated H3K27 were reduced in equivalent knock-outs carried out in mouse ESC lines;[192] these data again illustrate a direct link between PRC2 and H3K27 methylation. This is not surprising given that PRC2 is the only enzyme complex yet known to catalyze this reaction in mammals.[191] Importantly, experimental evidence shows that the regulation of cellular differentiation and development is supported by PRC1-mediated H2AK119 monoubiquitylation.[224,225] This enzymatic activity is stimulated in part by H3K27me3 recognition (see previous section) but may also occur independently of PRC2.[218,226] Like knock-outs of PRC2 members, *Ring1B* knock-outs are lethal to embryos.[227] Furthermore, mouse ESCs lacking *Ring1B* show reduced levels of H2AK119 monoubiquitylation and impaired differentiation in culture.[226,228]

Given that PRC1 and PRC2 are tightly involved in determining whether cells remain pluripotent or become differentiated, it is not surprising that disturbances of this balance can lead to cancer development and metastasis.[224] Indeed, PRC2 members *EZH2*, *EED* and *SUZ12* have all been found to be disrupted in a variety of human cancers,[180,195,217] and particularly in haematological malignancies.[229–233] A longer discussion on *EZH2* alterations in cancer is provided in section 1.2.1.2 (which follows). PRC1 involvement in cancer pathogenesis is best exemplified by the deregulation of BMI-1 (leukemia viral BMI-1 proto-oncogene, polycomb ring finger, OMIM *164831) which has been observed in multiple cancers,[180,184,224,234] and sometimes in combination with overexpression of EZH2 (see section 1.2.1.2.1).

Finally, another important gene silencing activity that involves PRC2 (again supported by PRC1) is X-inactivation.[225,226] "X-inactivation" refers to the process by which one X

chromosome is silenced in females to establish normal dosage of X-linked genes, and through which this silencing is maintained during subsequent cell divisions.[235,236] X-inactivation begins when a non-coding RNA called *XIST* coats the inactive X chromosome (Xi) to silence it;[235,236] the accompanying increase in *XIST* RNA expression rapidly recruits PRC2 for the establishment of H3K27me3 marks.[235,236] This H3K27 methylation is important during the initiation of X inactivation, but does not directly assist in the maintenance of Xi silencing.[235,236] It has been proposed that PRC2-mediated H3K27me3 is required to stabilize the Xi chromatin structure, possibly by recruiting other repressive machinery, which in turn would help maintain Xi silencing throughout cell divisions.[236] A similar mechanism pairing non-coding RNAs and PRC2 for regional silencing has been suggested in imprinting.[237,238]

#### 1.2.1.2 *EZH2* alterations in cancer

*EZH2* was first cloned from humans in 1996 as part of a project designed to identify genes on chromosome 21 that might contribute to the phenotype of Down syndrome;[185] it showed strong homology to the *Drosophila* enhancer of zeste protein encoded by the *E(z)* gene and so was named *EZH2* (as the name *EZH1* had already been given to another gene encoding a similar protein). Later on, Cardoso *et al*.[239] showed that this sequence on chromosome 21 was in fact a pseudogene, and that the functional *EZH2* gene actually maps to chromosome 7q36.1. They also showed that *EZH2* is composed of 20 exons, with an open reading frame spanning from exon 2 to 20 (as shown in Figure 1-1).[239] The 7q36 region had been previously associated with myeloid disorders, so the authors proposed that EZH2 might be involved in cancer pathogenesis.[239]

*EZH2* and/or EZH2 expression have since been shown to be altered in cancer, through diverse mechanisms that may be dependent on cancer type.[240] Point mutations, copy number variants and gene expression changes have all been observed across numerous cancers types, and especially in haematological malignancies (see Appendix A for the distribution of *EZH2* alterations across tissues according to the COSMIC database).

##### 1.2.1.2.1 Overexpression and/or amplification

Aberrant expression of EZH2 was first observed in Hodgkin and non-Hodgkin lymphomas, in combination with BMI-1 expression; the two proteins (which usually assemble into PRC2 and PRC1 respectively) had rarely been detected in the same cells and were thought (at that time) to

be mutually exclusive.[241,242] These studies represent the first *in vivo* evidence for a role of EZH2 expression in cancer. Soon after, in 2002, overexpression of EZH2 was described in prostate cancer.[243] Not only was this overexpression (of both transcript and protein) associated with metastatic disease, higher concentrations of EZH2 directly correlated with a poorer prognosis.[243] Similar findings followed in breast cancer.[244] Since then, elevated expression of EZH2 has been observed in a multitude of cancers such as bladder,[245] ovarian,[246] gastric,[247] liver,[248] skin,[249] and others.[195] *In vitro* experiments using small interfering RNAs targeted against EZH2 in cancer cell lines showed successful inhibition of cell proliferation, suggesting that EZH2 may be a good target for chemotherapy.[243,250]

Interestingly, Xu *et al*.[251] observed that overexpressed EZH2 promoted progression of prostate cancer to a lethal metastatic (castration-resistant) state in a PRC2-independent manner.[252] Through ChIP-seq experiments (chromatin immunoprecipitation combined with next-generation DNA sequencing), the authors discovered that some EZH2-bound sites lacked nearby H3K27me3 marks, thus they named these "solo" sites.[251] Silencing of SUZ12, essential within PRC2, did not affect EZH2 binding at these "solo" sites, but did reduce binding at sites near silencing marks, suggesting that EZH2 binding at "solo" sites was not dependent on PRC2 assembly, whereas binding at other sites did require PRC2 assembly.[251] Further, these "solo" sites were actually enriched with activating histone marks (H3K4me2/3),[251] suggesting that EZH2 can act as an activator rather than repressor under certain circumstances, an insight that could be explored for new therapeutic inhibitory strategies.[251] This activator role is thought to be dependent on phosphorylation of EZH2 by AKT,[208] which shifts the HMT specificity towards the androgen receptor (AR) and AR-associated proteins by promoting the direct association of EZH2 with AR-containing complexes.[251]

#### 1.2.1.2.2 Somatic missense variants

High-throughput sequencing technologies have uncovered widespread somatic variation of *EZH2* in cancer, with the highest number of mutations reported in haematological malignancies. The first observation was made in 2010 by Morin *et al*., who identified recurrent missense mutations altering tyrosine residue 646 (or 641 in the shorter EZH2 isoform,[253] as reported in the original publication) in follicular and diffuse large B-cell lymphomas of germinal-center origin.[254] Since then, independent studies have confirmed that this residue is a hotspot for

mutations in diffuse large B-cell lymphomas[255] and, at lower frequencies, in other cancers such as melanomas.[256] Functional studies originally suggested a loss-of-function mechanism for mutations affecting tyrosine 646,[254] but further studies have now proven that these mutations are activating and lead to increased trimethylation of H3K27.[190,257] Only two other residues, alanine 682 and alanine 692 (corresponding to residues 677 and 687 in the shorter EZH2 isoform), have been linked to gain-of-function mutations (see Chapter 3, section 3.1.1).[258,259] Malignancies with these mutations would likely benefit from EZH2 inhibitors,[250,260–263] similarly to malignancies with overexpressed EZH2 as discussed in the previous section.

Somatic mutations of *EZH2* have also been described in T-cell acute lymphoblastic leukemias,[232] and myeloid disorders.[264,265] These mutations are mostly missense, but nonsense and stop codon mutations, as well as deletions, have also been observed, with some being homozygous. Furthermore, they are distributed across the gene, and thus predicted to result in loss-of-function with regards to PRC2-mediated HMT activity.[266] These observations suggest that a precise amount of EZH2-regulated H3K27 methylation is required for normal cell development and proliferation, and that the exact mechanism leading to EZH2 dysregulation in each cancer type should be determined before initiating treatment with the EZH2 inhibitors that are currently being developed and tested.[260–263,266,267]

#### 1.2.1.3 The normal spectrum of genetic variation in *EZH2*

The human genome is highly variable;[268] thus, as expected, normal variation is observed within the *EZH2* coding region. This variation is summarized in Appendix B. Interestingly, there is only one truly common variant described, p.(Asp185His), which is found in 8% of the population according to dbSNP (rs2302427). All other variants are described in very few or even single individuals. This suggests that *EZH2* is highly intolerant to sequence variation, which was expected due to its crucial role in mammalian development.

The Exome Aggregation Consortium (ExAc) created z-scores to assess the tolerance of each gene to variation.[269] Any score greater than zero indicates intolerance, with increasing scores reflecting increasing intolerance. The z-score for missense mutations in *EZH2* is 5.45, which is very high, and consistent with the reduced variation observed in this gene (only 77 missense variants observed compared to 255 expected). ExAc also predicts *EZH2* to be highly intolerant to loss-of-function mutations (with zero mutations at essential splice sites, and zero stop-gains

observed vs. 32 expected, for a pLI score of 1.00). For comparison purposes, the z-score for missense mutations in *NSD1* is 2.38 (with 693 variants observed vs. 834 expected), which could explain why Sotos syndrome is more common than Weaver syndrome.

### 1.2.2 NSD1

Human *NSD1* was first isolated in 2001 due to its involvement in acute myeloid leukemia (AML). By investigating the breakpoints of the recurrent t(5;11)(q35;p15.5) translocation observed in AML, usually associated with del(5q), the authors isolated a chimeric messenger RNA resulting from the fusion of *NUP98* to an unknown gene at 5q35.[270] Using BLAST, they found that this gene was highly homologous to the mouse *Nsd1* gene and thus it was recognised as human *NSD1*.[270] The gene was also characterized by an independent group who showed that *NSD1* is composed of 23 exons, with an open reading frame spanning from exon 2 to 23.[271] The *NUP98-NSD1* fusion gene has since been described in approximately 16% of pediatric AML patients[272] and 2% of adult AML patients,[272,273] and is associated with poor prognosis.[272–274]

A lot of parallels may be drawn between NSD1 and EZH2. First of all, NSD1 is also a SET-domain containing histone methyltransferase, but with specificity towards lysine residue 36 of histone 3 (H3K36).[275] NSD1 can mono- and di-methylate H3K36, while SETD2 carries out the trimethylation step;[276,277] H3K36 methylation is associated with active gene transcription,[276,277] and antagonizes PRC2-mediated H3K27 methylation (as mentioned earlier). Second, *HOX* gene expression was altered in *NUP98-NSD1*-positive AML,[272] with higher expression of *HOXA* and *HOXB* genes, supporting a role for NSD1 in the regulation of homeotic gene expression. Wang *et al*. hypothesized that this overexpression could result from NSD1-mediated H3K36 methylation interfering with EZH2-mediated repression of these genes, consistent with the antagonistic methylation of H3K36 and H3K27 on the same H3 tail.[278] In addition, the *Nsd1* mouse knock-out is embryonic lethal, which proves that NSD1 also plays an important role in mammalian development.[275] Finally, alterations of the *NSD1* gene have been detected in various cancers: in addition to the fusion protein discussed above that is observed in AML, somatic mutations and/or copy number variants have been described in a variety of cancers such as lung,[279] prostate,[280] neuroblastomas and glioblastomas.[281]

## 1.3 Doctoral thesis framework

### 1.3.1 Rationale and hypothesis

My main research question is the following: which genes and mechanisms cause rare overgrowth syndromes that are phenotypically similar to Weaver syndrome?

For the work presented here, I hypothesized that genes involved in Weaver-like overgrowth syndromes are other members of the PRC2 complex, and/or other chromatin regulators. I used two main tools to investigate my hypothesis: detailed phenotyping of patients to characterize different overgrowth syndromes, and next generation-sequencing to enable new gene discoveries.

### 1.3.2 Objectives and research plan summary

In order to address the hypothesis above, I divided my research plan into three separate aims as described below. Each aim corresponds to a separate chapter (Chapters 2-4).

Aim 1: Screen a cohort of patients with Weaver-like features for mutations in *EZH2* and *NSD1*.

Aim 1a: Gather detailed phenotypic information.

Aim 1b: Screen for point mutations and small indels in the coding region of *EZH2*.

Aim 1c: When no mutations are found, screen for point mutations and small indels in the coding region of *NSD1*.

Aim 2: Use *in vitro* functional studies to measure the effect of *EZH2* mutations on PRC2-mediated histone methyltransferase activity, and determine whether this effect correlates with the severity of the Weaver syndrome phenotype.

Aim 3: Identify the underlying genetic defect in patients with Weaver-like features that do not have mutations in *EZH2* nor *NSD1* via exome sequencing, and use detailed phenotypic data (collected in Aim 1a) to aid in variant interpretation.

# Chapter 2: Detailed phenotyping and genotyping of patients clinically suspected of having Weaver syndrome

## 2.1 Background

Following the discovery that rare *de novo* mutations in *EZH2* cause Weaver syndrome,[56] the Gibson laboratory began to establish an overgrowth cohort. Patients were recruited at the BC Children's Hospital (BCCH) Medical Genetics Clinic, through physicians from other hospitals referring their patients to us, or from families that contacted us directly. The objective was to offer research-grade molecular testing to these patients, while collecting detailed phenotypic and genetic information on each patient to further understand the phenotypic and mutational spectrums of Weaver and Weaver-like overgrowth syndromes.

## 2.2 Phenotyping

### 2.2.1 Methods

#### 2.2.1.1 Inclusion criteria

Patients who had a clinical diagnosis or clinical suspicion of Weaver syndrome were accepted into this overgrowth study. Inclusion criteria included the following:

1- generalized overgrowth (several height and/or weight measurements above the 85$^{th}$ percentile or +2 standard deviations (S.D.) for age according to the CDC clinical growth charts, available at http://www.cdc.gov/growthcharts/clinical_charts.htm);

2- macrocephaly (several head circumference measurements above the 85$^{th}$ percentile or +2 S.D. for age according to the CDC clinical growth charts);

3- accelerated osseous maturation determined by X-rays (over +1 S.D.);

4- Weaver-like facial dysmorphism (mentioned in clinical letters or determined through photographs);

5- intellectual disability and/or developmental delay;

6- early cancer development (under 30 years of age);

7- congenital abnormalities;

8- negative microarray results (clinical-grade).

Further, individuals with overgrowth features and positive *EZH2* or *NSD1* testing (with pathogenic variants, possibly pathogenic variants or variants of unknown significance identified in an external laboratory) were also enrolled.

#### 2.2.1.2 Consenting

Participating families provided informed consent specifically to investigate the cause of their disease. The possibility of uncovering informative variants unrelated to the main phenotype but medically actionable (known as incidental findings) was discussed. Consent was provided directly through our laboratory or indirectly through their referring physician or genetic counsellor, and this study was approved by the joint Clinical Research Ethics Board at UBC and BCCH as described before. After consent was obtained, study participants were assigned a Gibson laboratory study number, as well as an overgrowth-specific study number; the latter numbering system is used throughout this thesis.

#### 2.2.1.3 Collection of phenotypic information

Clinical data were collected via mail, secure email or fax. Documents were sent by the physicians who referred the patients, by the families themselves, or by hospital staff when families provided us with a release of information form. Documents collected included: official health records, clinical letters, photographs at various ages, growth curves, previous genetic test results, X-rays and magnetic resonance imaging (MRI) scans. The number of documents collected and the level of detail in such documents varied greatly between patients. These data were compiled on an extensive table using Microsoft Excel.

#### 2.2.1.4 Determination of growth percentiles

Growth percentiles were calculated using freely available online tools, most of which are based on the CDC growth charts (Centres for Disease Control and Prevention, United States).

Percentiles corresponding to birth weight, height and head circumference were determined using only the Fenton 2013 growth charts [282] for preterm babies (≤ 37 weeks gestational age), or using both the Fenton 2013 and the 2006 WHO growth charts (for infants from 0 to 24 months of age) for babies born after 37 weeks gestational age. These tools can be accessed at http://peditools.org/fenton2013/ and http://peditools.org/growthwho/.

Percentiles corresponding to weight, height and head circumference of infants under 36 months (i.e. 3 years) of age were determined using the 2000 CDC growth charts (for infants from 0 to 36 months of age), accessible at http://peditools.org/growthinfant/.

Percentiles corresponding to weight and height for individuals between 3 and 20 years of age were determined using the 2000 CDC growth charts (for children and adolescents from 2 to 20 years of age), accessible at http://peditools.org/growthpedi/index.php**.** For individuals over 20 years of age, adult percentiles (at 20 years) were given. Percentiles corresponding to head circumference measurements on all individuals over 3 years of age were estimated from the 2000 CDC growth charts via http://www.simulconsult.com/resources/measurement.html?type=head.

### 2.2.2 Description of cohort

#### 2.2.2.1 Complete overgrowth cohort

As of June 1st 2016, 66 individuals from 64 independent families had been recruited to the overgrowth study. Within this cohort, only the proband was deemed affected in 61 out of 64 families; thus the majority of cases are sporadic, which is consistent with what has been reported in the literature (see Chapter 1). To the best of our knowledge, only probands 1-4 had been published prior to the work presented in this thesis (see section 2.3.2.2.1 for details).

In our cohort, there are 40 males (61%) and 26 females (39%), ranging from age 1 year and 6 months to 44 years and 8 months (mean age: 12 years and 7 months; median: 10 years and 9 months). Specific ancestry information is not available for 10/66 patients, but the majority (at least 41/66, or 62%) are reported to be of European descent. Among the other 15 individuals, three are African American, three are from the Middle-East, four are "latino", one is Asian, and four have mixed ethnicity backgrounds. Most patients were recruited from Canada (N=23) or the United States (N=21), but others were recruited from the following countries: Argentina (1), Australia (5), Chile (1), France (1), Italy (1), Mexico (1), New Zealand (1), Norway (2), Portugal (4), South Africa (1), Spain (1), Turkey (2), Ukraine (1). This international recruitment across 15 different countries was necessary for collecting sufficient patients with rare Weaver-like overgrowth.

The prevalence of each phenotypic trait is described in Table 2-1 (expanded from Gibson *et al.*[56]). As expected, the features most commonly observed within our cohort (at >80% frequency) are unspecific and shared between a large number of genetic disorders; these include speech

delay (93%), intellectual disability (90%), developmental delay (84%) and poor fine motor coordination (83%). Next, reflecting our targeted recruitment, some features characteristic of overgrowth syndromes were also common (with 70-80% frequency), including: accelerated osseous maturation (79%), prominent forehead (78%), macrocephaly (78%), tall stature (77%), excessive growth of prenatal onset (74%), ocular hypertelorism (72%), and poor balance (72%). A heterogeneous presentation of behavioural and/or mental disorders was also common in this cohort (at 74% frequency, excluding autism spectrum disorders). These behavioural disorders and poor balance are likely to be under-ascertained due to some of our study participants being too young to assess these features properly.

Other traits that occur at relatively high frequency in our cohort are: hypotonia, brain and cardiovascular abnormalities, hoarse low-pitched cry, rounded face in early years, and micro/retrognathia. Cancer prevalence in cohort members is 6% (as ascertained by June 1st 2016) for a total of four cases (three with neuroblastomas and one with leukemia), which is comparable to that observed among overgrowth syndromes (see Chapter 1).

#### 2.2.2.2 Weaver-like cohort

After collecting more detailed clinical information, individuals with overgrowth were re-classified based on whether their phenotype was consistent with a Weaver- or Sotos-like syndrome. For this purpose, individuals were expected to have tall stature and/or macrocephaly, together with at least one of the following:

1- accelerated osseous maturation;

2- Weaver-like dysmorphic features (with a combination of at least three of the following: rounded face in early years, prominent forehead, large ears, hypertelorism, down-slanting palpebral fissures, micro/retrognathia or "stuck-on" chin);

3- early cancer development;

4- other major congenital abnormalities (such as those affecting the brain or heart);

5- umbilical and/or inguinal hernias.

Based on these criteria alone (using the Excel database), only three individuals were excluded, mainly because phenotypic information was insufficient to make an informed decision about inclusion/exclusion for a large number of study participants. Next, we re-analyzed in depth the patients' health records, family history, and photographs, to look for consistency with a

Weaver- or Sotos-like phenotype. Through this more subjective method, a further 14 individuals were excluded. Lastly, two individuals who were found externally to have an alternative molecular diagnosis while our investigations were ongoing were also excluded. Phenotypic analysis was repeated without these 19 individuals (new N=47) to determine whether selecting for a specific phenotype would shift the overall frequencies of Weaver- and Sotos-specific phenotypic features (see Table 2-1).

Within the "Weaver-like" cohort, the most common features still include speech delay and intellectual disability at a frequency >90%. Additionally, some overgrowth-related features appear to have increased in frequency, crossing the 80% threshold, which suggests that the criteria used to define this sub-cohort were effective; these features include: tall stature (increasing from 77 to 87%), excessive growth of prenatal onset (74 to 81%) and macrocephaly (78 to 83%).

Other relative increases in frequency to note include several Weaver-like features such as: umbilical hernias (44 to 63%), hypoplastic/supernumerary nipples (40 to 53%), micro/retrognathia (60 to 73%), large ears (42 to 51%), down-slanted palpebral fissures (53 to 61%), ocular hypertelorism (72 to 80%), hoarse low-pitched cry (61 to 67%), and others. Further, although the number of cases remained the same (N=4), the overall cancer frequency increased from 6 to 8.5%, which is closer to the reported range for Weaver syndrome (9.5-10.5%, see Chapter 1).

To note, the most substantial decrease in frequency was for congenital heart defects, and particularly the prevalence of septal defects which decreased from 69 to 15%.

| | Prevalence in overgrowth cohort (N = 66) * | Prevalence in patients with Weaver-like features (N = 47) * |
|---|---|---|
| **General** | | |
| Age range (years and months) | 1y6m - 44y8m | 1y6m - 44y8m |
| Sex distribution (males; females) | 40 (61%); 26 (39%) | 29 (62%); 18 (38%) |
| Gestational age at delivery (weeks and days) | 29w3d - 42w | 32w - 42w |
| **Growth features** | | |
| Birth weight (percentile range) | $<5^{th}$ - $>99^{th}$ | $<5^{th}$ - $>99^{th}$ |
| Birth length (percentile range) | $32^{nd}$ - $>99^{th}$ | $32^{nd}$ - $>99^{th}$ |
| Birth head circumference (percentile range) | 10th - $>99^{th}$ | 10th - $>99^{th}$ |
| Excessive growth of PREnatal onset | 37/50 (74%) | 29/36 (81%) |
| Excessive growth of POSTnatal onset | 9/50 (18%) | 7/36 (19%) |
| Tall stature | 48/62 (77%) | 39/45 (87%) |
| Obesity | 22/52 (42%) | 15/35 (43%) |
| Accelerated osseous maturation | 34/43 (79%) | 29/37 (78%) |
| **Neurological features** | | |
| Hypertonia | 9/49 (18%) | 8/35 (23%) |
| Hypotonia | 31/51 (61%) | 19/34 (56%) |
| Hoarse low-pitched cry | 14/23 (61%) | 12/18 (67%) |
| Intellectual disability | 36/40 (90%) | 27/30 (90%) |
| Excessive appetite | 16/42 (38%) | 11/29 (38%) |
| Developmental Delay | 48/57 (84%) | 31/39 (80%) |
| Speech Delay | 49/53 (93%) | 34/37 (92%) |
| Autism spectrum disorder | 15/50 (30%) | 7/35 (20%) |
| Other behavioural and/or mental disorders | 37/50 (74%) | 25/35 (71%) |
| Ventriculomegaly | 6/37 (16%) | 5/26 (19%) |
| Delayed myelination | 3/32 (9%) | 1/22 (5%) |
| Cerebellar hypoplasia | 2/36 (6%) | 1/25 (4%) |
| Seizures | 14/50 (28%) | 11/34 (32%) |
| Polymicrogyria | 1/35 (3%) | 1/24 (4%) |
| Pachygyria | 2/36 (6%) | 2/25 (8%) |
| Other brain abnormalities | 26/44 (59%) | 14/29 (48%) |
| Poor fine motor coordination | 40/48 (83%) | 30/34 (88%) |
| Poor balance/gravitational insecurity | 28/39 (72%) | 23/29 (79%) |
| **Craniofacial** | | |
| Macrocephaly | 49/63 (78%) | 38/46 (83%) |
| Rounded head/face (at least in early years) | 35/58 (60%) | 26/45 (58%) |
| Large bifrontal diameter | 30/53 (57%) | 23/41 (56%) |
| Prominent forehead | 46/59 (78%) | 33/44 (75%) |
| Flat occiput | 18/35 (51%) | 15/27 (56%) |
| Large ears ** | 26/62 (42%) | 23/45 (51%) |
| Otitis media (recurrent) | 9/62 (15%) | 7/36 (19%) |
| Ocular hypertelorism | 39/54 (72%) | 31/39 (80%) |
| Strabismus | 6/27 (22%) | 3/16 (19%) |
| Myopia | 9/38 (24%) | 4/36 (11%) |
| Other eye abnormalities | 15/38 (40%) | 13/36 (36%) |
| Down slanted palpebral fissures | 31/59 (53%) | 26/43 (61%) |
| Full/thick eyebrows | 7/54 (13%) | 7/42 (17%) |
| Sparse eyebrows | 15/54 (28%) | 11/42 (26%) |
| Long philtrum | 25/57 (44%) | 19/43 (44%) |
| Broad nose | 5/53 (9%) | 3/41 (7%) |
| Wide nasal root | 28/53 (53%) | 21/39 (54%) |
| High arched palate | 17/36 (47%) | 13/22 (59%) |

| | Prevalence in overgrowth cohort (N = 66) * | Prevalence in patients with Weaver-like features (N = 47) * |
|---|---|---|
| Prominent chin/jaw | 26/55 (47%) | 21/42 (50%) |
| Micro/retrognathia | 33/55 (60%) | 29/40 (73%) |
| Rosy cheeks or malar flushing | 17/51 (33%) | 15/40 (38%) |
| **Cardiovascular** | | |
| Patent ductus arteriosus | 3/18 (17%) | 2/14 (14%) |
| Septal defects (one atrial, others ventricular) | 11/16 (69%) | 2/13 (15%) |
| Respiratory problems | 18/38 (47%) | 12/25 (48%) |
| **Limbs** | | |
| Limited elbow & knee extension in early life | 12/43 (28%) | 10/30 (33%) |
| Limited elbow & knee extension after puberty | 6/11 (55%) | 5/9 (56%) |
| Widened distal femurs and ulnas | 3/9 (33%) | 2/8 (25%) |
| **Hands** | | |
| Large hands ** | 19/36 (53%) | 16/26 (62%) |
| Long slender fingers | 7/34 (21%) | 4/27 (15%) |
| Prominent digit pads | 18/45 (40%) | 16/33 (49%) |
| Single transverse palmar crease | 9/40 (23%) | 6/28 (21%) |
| Camptodactyly | 8/47 (17%) | 6/35 (17%) |
| Clinodactyly | 6/45 (13%) | 4/34 (12%) |
| Broad thumbs | 12/43 (28%) | 8/31 (26%) |
| Thin, deep-set nails | 12/43 (28%) | 11/32 (34%) |
| **Feet** | | |
| Large feet ** | 14/32 (44%) | 11/21 (52%) |
| Clinodactyly (usually 4/5) | 13/50 (26%) | 10/37 (27%) |
| Syndactyly (2/3) | 3/49 (6%) | 3/37 (8%) |
| Talipes equinovarus | 6/40 (15%) | 5/28 (18%) |
| Short fourth metatarsals | 4/39 (10%) | 3/28 (11%) |
| Hind foot valgus | 10/40 (25%) | 8/29 (28%) |
| **Skin** | | |
| Excessive loose skin | 13/55 (24%) | 10/39 (26%) |
| Hypoplastic/supernumerary nipples | 18/45 (40%) | 17/32 (53%) |
| Thin hair | 10/47 (21%) | 6/33 (18%) |
| Increased pigmented nevi or other marks | 29/50 (58%) | 19/34 (56%) |
| **Connective tissue** | | |
| Umbilical hernia | 16/36 (44%) | 15/24 (63%) |
| Inguinal hernia | 4/35 (11%) | 3/23 (13%) |
| Diastasis recti | 6/42 (14%) | 5/29 (17%) |
| Scoliosis | 16/44 (36%) | 12/31 (39%) |
| Kyphosis | 4/46 (9%) | 3/34 (9%) |
| **Endocrine** | | |
| Hypothyroidism | 4/35 (11%) | 2/25 (8%) |
| Growth hormone deficiency | 2/26 (8%) | 2/20 (10%) |
| Hypoglycemia | 8/33 (24%) | 7/26 (27%) |
| **Neoplasia** | total 4/66= 6% | total 4/47= 8.5% |
| Neuroblastoma | 3/44 (7%) | 3/33 (9%) |
| Leukemia | 1/46 (2%) | 1/35 (3%) |
| Lymphoma | 0/46 (0%) | 0/35 (0%) |
| **Other** | | |
| Cryptorchidism | 6/21 (29%) | 5/13 (39%) |
| Early puberty | 7/21 (33%) | 3/13 (23%) |

**Table 2-1: Prevalence of each phenotypic trait in our overgrowth and Weaver-like cohorts.**

* Not every feature was described for every patient, thus only informative numbers and percentages are given for each feature. Further, this table refers to presence or absence of each feature in each patient and does not reflect prominence of the feature ("present" included mild to severe/very prominent). Characteristics that were described as present then resolved at a later date were also counted as present. Unspecific or unclear descriptions were considered non-informative. ** Not specified if proportional or not to generalized overgrowth.

## 2.3 Genotyping

### 2.3.1 DNA extraction and quality control

#### 2.3.1.1 DNA extraction from EDTA-anticoagulated blood

Blood received in tubes supplemented with EDTA (ethylene diamine tetraacetic acid; purple top) was processed using the E.Z.N.A. Blood DNA Maxi Kit from Omega Bio-tek (#D2492-03). Manufacturer's instructions were followed, except that columns were equilibrated with 3M sodium hydroxide before use (incubation for 5 mins followed by centrifugation at high speed for 5 mins), and that the final elution was done with 500 µl of Elution Buffer heated to 70°C.

Due to persistency of contaminants in our samples, DNA extraction from blood was followed by DNA precipitation using 3M sodium acetate ($1/10^{th}$ of the sample volume) and 1 ml of 100% ethanol. After mixing by inversion (10 times), each tube was incubated overnight at -20°C. High speed centrifugation (at 4°C) was used to separate the DNA pellet from the supernatant, which was discarded. Pellets were washed with cold 70% ethanol then left to dry at room temperature for 2-4 hours. DNA was re-suspended in 200 µl of Elution Buffer and incubated at room temperature for at least 1 hour before proceeding with quality control.

#### 2.3.1.2 DNA extraction from saliva

Saliva samples were collected in Oragene-DNA kits from DNA Genotek (OG-500 or OG-575 for assisted collection), which were mailed out (pre-labelled) to referring physicians or families directly. These kits were chosen because they are easy to use and come with a preservative that renders the sample stable at room temperature for years, thus simplifying mailing to and from remote locations, including others countries.

DNA extraction from saliva was carried out as per DNA Genotek's instructions using their prepIT-L2P buffer (#PT-L2P) provided with the Oragene-DNA collection kits. After the DNA

was pelleted and washed with 70% ethanol, a drying step was added to allow for the residual ethanol to evaporate by leaving the tubes open at room temperature for 10 mins. DNA was then re-suspended in 200 µl of Elution Buffer and incubated at room temperature overnight, with occasional vortexing. Because this protocol did not include RNase treatment, this was carried out immediately after DNA extraction.

For RNase treatment, RNase A (solution at 10 µg/µl from Omega Bio-tek) was added to the DNA for a final concentration of 10 µg/ml (so 0.5 µl of RNase solution added to 500µl of DNA). Incubation was carried out at 37°C for 30 mins using a dry heating block, then followed by DNA precipitation using 3M sodium acetate as described above, and quality control.

#### 2.3.1.3 DNA extraction from nail clippings

DNA extraction from nail clippings was carried out using the QIAamp DNA Investigator Kit (QIAGEN #56504) as per manufacturer's instructions. The final elution was done with 36 µl of distilled water. DNA precipitation using 3M sodium acetate as described above was not carried out due to the low yield of this extraction procedure; quality control was carried out immediately following extraction.

#### 2.3.1.4 Quality control

Quality control of DNA samples extracted in our lab or received from our collaborators was carried out in two steps. First, the NanoDrop 2000c UV-Vis spectrophotometer (Thermo Scientific) was used to quantify the DNA (at 260 nm) and to calculate the sample purity ratios (260/280 nm and 260/230 nm). Based on these values, a volume corresponding to 250-300 ng of DNA was mixed with 6X DNA loading dye (Thermo Scientific #R0611) and loaded onto a 1% agarose gel prepared with 1X TBE buffer and SYBR Safe DNA gel stain (Invitrogen #S33102). The GeneRuler 1 kb Plus DNA ladder from Thermo Scientific (#SM1333) was used as a size and concentration control marker. Gels were imaged using the UVP BioSpectrum 310 Imaging System. A single clean band on the gel was indicative of high quality DNA that could be used for sequencing.

### 2.3.2 Sanger sequencing of *EZH2* and *NSD1*

#### 2.3.2.1 Methods

##### 2.3.2.1.1 Primer design for PCR and Sanger sequencing of *EZH2*

*EZH2* is made up of 20 exons, of which 19 are coding (2-20),[239] as described in Figure 2-1. NCBI Reference sequence NM_004456.4 is available from GenBank; all work described here as well as nomenclature of all sequence variants in *EZH2* are based on this sequence. This corresponds to the longest isoform of the EZH2 protein (751 amino acids);[253] the resultant protein sequence, NP_004447.2, is available from GenPept. Further protein domain information was extracted from UniProtKB (Q15910), available at http://www.uniprot.org**.**

Primers were designed to amplify each coding exon in a separate reaction. Primer sequences were obtained from Gibson *et al.*[56] and can be found in Appendix C.1. All primer pairs were confirmed to be specific to the *EZH2* gene on chromosome 7 (NC_000007.13: 148504464-148581441) rather than the pseudogene on chromosome 21 (NC_000021.8: 36971977-36972553) using the BLAST function on NCBI at http://blast.ncbi.nlm.nih.gov/Blast.cgi (GRCh37.p13 assembly). Primers were also confirmed to be of similar melting temperatures (reflecting length and GC content) and not to be self-binding (to avoid hairpin formation and reduced efficiency) at http://biotools.nubic.northwestern.edu/OligoCalc.html**.**

These specific primers (custom DNA oligos) were ordered from Integrated DNA Technologies and re-hydrated with distilled water to 100 μM stock solutions as per manufacturer's instructions. An aliquot of each primer stock was diluted to a 5 μM single primer solution for sequencing. A 5 μM primer pair mix (with 5 μM of forward and 5 μM of reverse primer) was also prepared for each coding exon for PCR amplification.

After sequencing of several patient samples, it was found that two common polymorphisms in exon 20 often interfered with the Sanger sequencing results, thus internal sequencing primers were designed (20FFi and 20RRi). These primer sequences may also be found in Appendix C.1.

##### 2.3.2.1.2 Primer design for PCR and Sanger sequencing of *NSD1*

*NSD1* is made up of 23 exons, of which 22 are coding (2-23),[271] as described in Figure 2-4. NCBI Reference sequence NM_022455.4 is available from GenBank; all work described here as well as nomenclature of sequence variants in *NSD1* are based on this sequence. The corresponding protein sequence, NP_071900.2, is 2696 amino acids long (see Figure 2-4) and is

available from GenPept. Further protein domain information was extracted from UniProtKB (Q96L73), available at http://www.uniprot.org**.** Primer sequences were obtained from Douglas *et al*.[283] and Rio et al.,[130] or designed using Primer-BLAST, a tool from NCBI that assists in designing primers while also using BLAST to confirm that these primers are specific to a particular locus and are not expected to bind to similar sequences elsewhere in the genome. This tool is accessible from the whole gene view of any particular gene, searchable at http://www.ncbi.nlm.nih.gov/gene/**.** Quality of primers was determined as before; primer sequences can be found in Appendix C.2. Primer preparation was also done as before.

#### 2.3.2.1.3 Sanger sequencing

After optimization using gradient PCR to determine the optimal annealing temperature for all primer pairs, PCR was performed on genomic DNA using a standard protocol (see Appendix D.1). PCR conditions for all *EZH2* primer pairs utilized on our thermocycler (Veriti 96-Well Thermal Cycler, Applied Biosystems) may be found in Appendix D.2, and PCR conditions for all *NSD1* primer pairs can be found in Appendix D.3. After PCR, 2 µl of each PCR product (representing each exon) were run on a 1.5% agarose gel (prepared as before) for quality control. Estimation of product concentration was done by comparing the intensity of our product bands to the bands produced by the GeneRuler 100 bp DNA ladder from Thermo Scientific (#SM0243). Gels were imaged as described previously.

PCR products equivalent to 70-100 ng (dependent on product size, where a minimum of 70 ng were added for products ≤ 300 bp) were transferred to new PCR tubes or plates and treated to remove residual primers and dNTPs. This PCR cleanup treatment was carried out with ExoSAP-IT (Affymetrix #78201) as per the BCCH Molecular Genetics Laboratory protocol: 2 µl of ExoSAP-IT mix per 8 µl of PCR product, treated at 37°C for 30 mins, followed by inactivation at 80°C for 15 mins. Alternatively, PCR cleanup enzymes from Thermo Scientific were used as follows: 0.5 µl Exonuclease I (20 U/µl, #EN0582) and 1 µl FastAP Thermosensitive Alkaline Phosphatase (1 U/µl, #EF0652) were mixed and added to 8 µl of PCR product, treated at 37°C for 30 mins, followed by inactivation at 85°C for 15 mins. When PCR plates were used, the plates were covered with Easy-Peel heat sealing foil (Thermo Scientific #AB-0745) and sealed

prior to incubation to avoid evaporation, using the heat sealer at the Centre for Molecular Medicine and Therapeutics (CCMT) sequencing core facility.

Following cleanup, tubes and/or plates were spun down to ensure that all of the treated product was at the bottom of the tubes; this centrifugation step also helped avoid cross-contamination of samples. Finally, 2 μl of each 5 μM primer solution were added to each reaction, and distilled water was added for a final volume of 15 μl. The mixtures of PCR products and primers were submitted for Sanger sequencing at the CMMT sequencing core facility; sequencing was carried out using their standard in-house protocol.

#### 2.3.2.1.4 Sequence analysis

Sanger sequences were analyzed using two different softwares. First, sequences were aligned to the reference for each exon, using CLC Sequence Viewer 6. This alignment allowed for detection of homozygous alterations and small indels (both in and out of frame). Some but not all heterozygous mutations were flagged in this alignment, thus we used Sequence Scanner v1.0 to visualize the actual Sanger peaks. The beginning and end of each exon were manually annotated, and Sanger peaks within the exon boundaries and near the intron-exon boundaries (5' and 3' of the exon) were scanned in detail to look for double peaks, which represent a heterozygous state at that particular locus. When variants were identified, we proceeded to variant interpretation and validation in parental samples (if applicable). When samples from clinically-unaffected siblings were available, these were also tested. Family relationships were interrogated using a "fingerprint" panel run on Sequenom at the Centre for Applied Neurogenetics (UBC). This assay provided the genotype call for 51 SNPs, of which only a subset were informative for each trio, so it did not definitely rule out non-relatedness but should be sufficient to rule out Mendelian errors and sample mix-up.

#### 2.3.2.1.5 Variant interpretation (for known disease genes)

Each variant and corresponding population frequencies were searched in the following public databases:

- dbSNP, database of short genetic variations, which includes the complete data from the 1000 Genomes project,[268] searchable at http://www.ncbi.nlm.nih.gov/projects/SNP/;

- Exome Variant Server (EVS), database from the NHLBI Exome Sequencing Project (ESP), the aim of which was to discover novel genes and mechanisms contributing to heart, lung and blood disorders by sequencing more than 200,000 richly-phenotyped individuals, searchable at http://evs.gs.washington.edu/EVS/;

- ExAc, database from the Exome Aggregation Consortium which includes genomic variation data from 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies,[269] searchable at http://exac.broadinstitute.org;

- ClinVar, which aggregates information about genomic variation and its relationship to human health, searchable at http://www.ncbi.nlm.nih.gov/clinvar/;

- DECIPHER, DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources, including primarily copy number variants but also sequence variants from the Deciphering Developmental Disorders (DDD-UK) project, searchable at https://decipher.sanger.ac.uk;

- LOVD, Leiden Open (source) Variation Database, whose purpose is to provide a flexible, freely available tool for Gene-centered collection and display of DNA variations, searchable at www.lovd.nl/EZH2 (v2.0) or http://databases.lovd.nl/shared/genes/EZH2 (v3.0) for *EZH2*, and at http://chromium.lovd.nl/LOVD2/home.php?select_db=NSD1 (v2.0) or http://databases.lovd.nl/shared/genes/NSD1 (v3.0) for *NSD1*;

- COSMIC, Catalogue Of Somatic Mutations In Cancer, relevant due to involvement of chromatin regulators such as EZH2 in cancer, searchable at http://cancer.sanger.ac.uk/cosmic.

The properties of these databases were reviewed by Johnston and Biesecker in 2013.[284] In 2015, the data from the ExAc and ClinVar databases became available within dbSNP, thus simplifying the search for variant information. The Human Gene Mutation Database (HGMD) was not used due to recent concerns that this database contains an excess of misclassified variants.[285]

Functional predictions for previously undescribed variants were done using the PROVEAN and SIFT tools, available at http://provean.jcvi.org/index.php (PROVEAN v1.1.3). Because these are *in silico* prediction methods, their scores were considered as supplementary information, whereas population frequencies and inheritance patterns were given more weighting.

Synonymous variants were classified as likely benign, or benign if inherited from an unaffected parent or reported at high frequency in the general population. Non-synonymous/

missense variants and small indels were classified as pathogenic when they were unique to the affected proband (confirmed *de novo* and not described in dbSNP, ExAc or the EVS), or when they were shared with other affected individuals (confirmed *de novo* in the proband and present in other individuals presenting with the same phenotype, as reported in the scientific literature, or in ClinVar, DECIPHER or LOVD databases, or segregating with the phenotype within the family). Variants that were described in dbSNP, ExAc or the EVS were not automatically excluded from analysis, but rather they were interpreted based on population frequency, as these databases currently contain both rare and common variants. Variants described with a Minor Allele Frequency (MAF) below 1% were considered to be rare and investigated further, and variants not described in any database where considered more likely to be pathogenic. Truncating variants were predicted to be likely pathogenic and interpreted further using the same guidelines as for missense variants. Variants affecting canonical splice sites (within five nucleotides of the intron/exon boundary) were generally classified as variants of unknown significance (VOUS); the exception is one variant found just one nucleotide outside the exon boundary and not described in any public database, which was classified as likely pathogenic. Missense variants described in public databases as being present in healthy individuals (with a MAF >1%) were classified as VOUS or likely benign depending on the population frequency and ancestry information available. All pathogenic variants identified in this study have been submitted to LOVD locus-specific databases (see above).

Importantly, this classification strategy is applicable to variants in genes that have been previously associated with disease, and only for patients whose phenotype is consistent with that previously associated with mutations in these genes. More evidence is required to classify novel variants in known disease genes found in patients with an unusual phenotype, and for variants in novel genes, as will be discussed in Chapter 4 (section 4.5.1).

#### 2.3.2.1.6 Reporting of Sanger sequencing results

Sequencing results were reported back to the referring physician, genetic counsellor or family member in the form of a research-grade sequencing report. An anonymized example of the *EZH2* report produced for proband 5, including the cover page, detailed report of the first coding exon (exon 2), detailed report of the exon containing a coding mutation (exon 18), and detailed report of two exons containing common intronic variants (exons 3 and 20), can be found

in Appendix E. For reports that listed variants of interest, we recommended validation of such variants by a CLIA-certified clinical laboratory so that the results could be used to guide clinical management. Further, informal genetic counselling was offered to families over phone and email, particularly when VOUS were reported. The limitations of our assay were made clear to everyone involved.

#### 2.3.2.2 Mutations and variants identified in *EZH2*

Of the 66 individuals (full overgrowth cohort), ten had been previously tested for variants in *EZH2*; two were found to be negative and the remaining eight were reported to have either pathogenic/likely pathogenic variants (5/8 including the three original patients from Gibson *et al*.[56]) or VOUS (3/8) in *EZH2*, which I validated independently. The entire coding region of *EZH2* was sequenced in 54 of the remaining 56 individuals. Two patients were not tested because their phenotype was more consistent with Sotos syndrome. Therefore, a total of 64 individuals were tested for variants in *EZH2*.

All variants identified or validated in the coding region of *EZH2*, or within 20 nucleotides either 5' or 3' of the exonic sequences, are summarized in Table 2-2. Other common variants located further away from the intron/exon boundaries that were observed upon Sanger analysis were included in the individual reports for completeness but are not included in Table 2-2.

##### 2.3.2.2.1 Rare variants in *EZH2*

Pathogenic mutations identified in *EZH2* are scattered throughout the gene, as illustrated in Figure 2-1. Patients with confirmed pathogenic mutations in *EZH2* were considered diagnosed and not investigated further; for description purposes, they were renamed as probands 1 to 11.

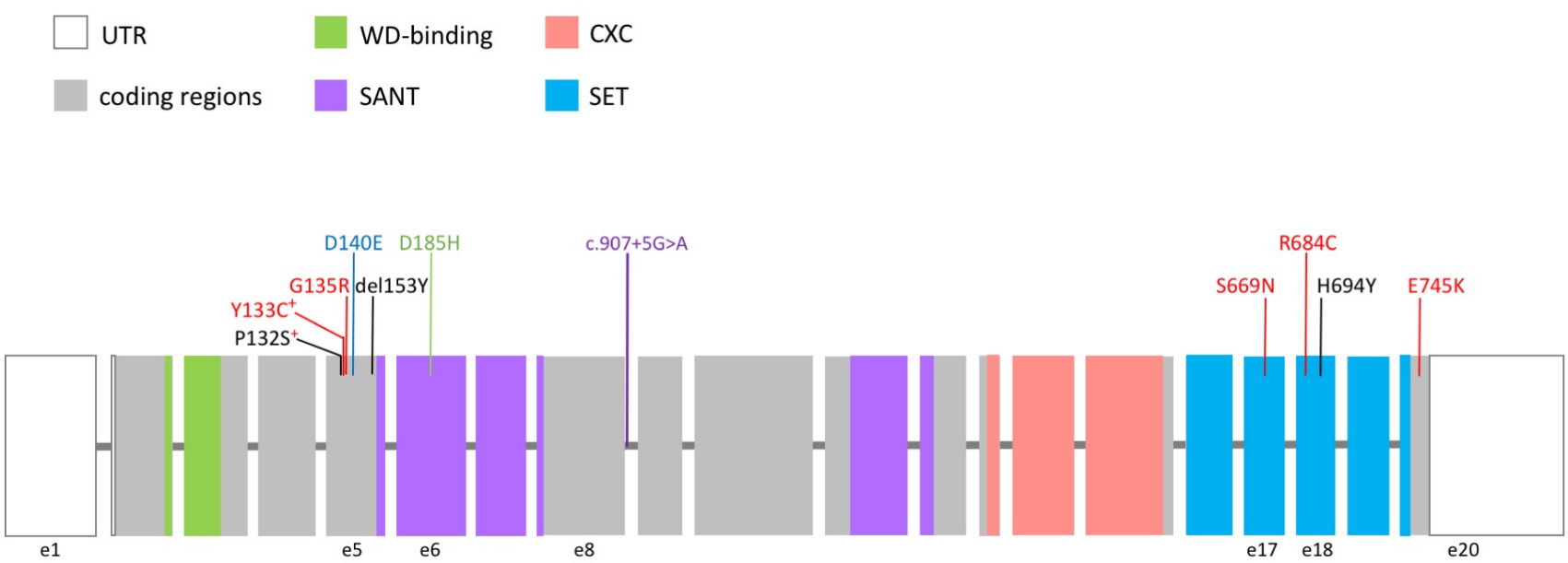



**Figure 2-1: Schematic representation of human EZH2.**

Human EZH2 is represented. Each rectangle represents one exon. Exon size is represented to scale, while intronic distances are not to scale. White (open) rectangles represent non-coding UTRs and grey rectangles represent coding exons (NM_004456.4). EZH2 protein contains 751 amino acids (NP_004447.2) and five recognizable domains (two SANT, and one each of WD-binding, CXC and SET), represented here in coloured rectangles according to NCBI (NP_004447.2) and UniProtKB/InterProt (Q15910-2) coordinates. Variants in black represent pathogenic mutations identified in the three original patients from Gibson *et al.*[56]; variants in red represent newly identified pathogenic mutations; variant in blue represents a VOUS; variant in purple represents a splice site VOUS; variant in green represents a likely benign common polymorphism observed in ten different individuals in our study.

[+] Mutations identified in two unrelated probands. e = exon; UTR = untranslated region; SANT = *Swi3, Ada2, N-Cor, and TFIIIB* domain; CXC = Cysteine rich motif; SET = *Su(var)3-9, E(z) and Trithorax* domain; VOUS = variant of unknown significance.

| Genomic position (chr 7, GRCh37.p13) | State | Position within gene structure | Predic-ted protein change | Presence in public databases and/or corresponding MAFs when applicable | | | | | | | Interpre-tation | Pre-valence among tested |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | dbSNP [a] | EVS (EA/All) | ExAc | Clin Var ID | DECI-PHER | EZH2 LOVD (v3.0) | COSMIC (COSM number) | | |
| 148543694 or 148543704 | het | flanking exon 3 (c.118-4delT) | n/a | rs3214332 (23.9%) or rs397889839 | nds | 71.3% | 210966 | nds | 000079 | 1735880 | benign | 54/54 |
| 148526910 | het | exon 5 (c.394C>T) | P132S | rs193921148 | nds | nds | 30200 | nds | 000058, 000054 | 133047 | pathogenic | 2/64 |
| 148526906 | het | exon 5 (c.398A>G) | Y133C | nds | nds | nds | nds | nds | 000055 | nds | pathogenic | 2/64 |
| 148526901 | het | exon 5 (c.403G>A) | G135R | nds | nds | nds | nds | nds | nds | 133044 | pathogenic | 1/64 |
| 148526884 | het | exon 5 (C.420T>A) | D140E | nds | nds | nds | nds | nds | nds | nds | likely pathogenic [c] | 1/64 |
| 148526845_ 148526847 | het | exon 5 (c.457_459delTAT) | del153Y | rs193921146 | nds | nds | nds | nds | 000052 | nds | pathogenic | 1/64 |
| 148525904 | het | exon 6 (c.553G>C) | D185H | rs2302427 (8%) | 8.4/6% | 7.9% | 134224 | nds | 000056 | 3762469 | likely benign | 10/58 |
| 148523541 | het | flanking exon 8 (c.907+5G>A) | n/a | rs368128494 | 0.03/ 0.02% | <0.1% | nds | nds | nds | nds | VOUS | 1/54 |
| 148514401 | het [b] | exon 11 (c.1323A>G) | E441= | nds | nds | nds | nds | nds | nds | nds | likely benign | 1/54 |
| 148511171 | het | exon 15 (c.1731G>A) | P577= | rs41277437 (0.06%) | 1.81/ 1.31% | 1.1% | 158577 | nds | nds | nds | likely benign | 2/54 |
| 148507448 | het | exon 17 (c.1991G>A) | S669N | nds | nds | nds | nds | nds | nds | nds | pathogenic | 1/64 |
| 148506462 | het | exon 18 (c.2050C>T) | R684C | rs587783626 | nds | nds | 158582 | 263342 | 000023 | 53005 | pathogenic | 1/64 |
| 148506432 | het | exon 18 (c.2080C>T) | H694Y | rs193921147 | nds | nds | 30199 | nds | 000053 | 53040 | pathogenic | 1/64 |
| 148506396 | het | flanking exon 18 (c.2110+6T>G) | n/a | rs41277434 (5%) | 5.42/ 4.27% | 6.2% | 137273 | nds | nds | 5020474 | likely benign | 4/54 |
| 148504761 | het | exon 20 (c.2233G>A) | E745K | rs397515548 | nds | nds | 65675 | nds | 000070 | 1087033 | pathogenic | 1/64 |
| 148504717 | het | flanking exon 20 (c.*+21delC) | n/a | rs3217095 (22.9%) | 31.3/ 23.1% | 69% | nds | nds | 000076 | nds | benign | 18/54 |
| | homo | | | | | | | | | | | 30/36 |

**Table 2-2: List of variants identified near or within the coding region of *EZH2*.**

[a] Minor allele frequencies (MAFs) provided for dbSNP correspond to the values given for the 1000 Genomes project.

[b] Somatic variant. [c] Not yet validated in parental samples.

Blue = exclusively previously identified variants that have been validated; het = heterozygous; homo = homozygous; n/a = non applicable; nds = not described; EA = European American population.

Probands 1, 2 and 3 were described previously.[56] The mutations identified in *EZH2* were c.457_459delTAT (p.Tyr153del), c.2080C>T (p.His694Tyr) and c.394C>T (p.Pro132Ser), respectively.

Proband 4 was originally published in 2001.[44] More detailed clinical features are described in Table 2-3. Using Sanger sequencing, we identified a c.394C>T (p.Pro132Ser) mutation in *EZH2*, an alteration previously described in proband 3 from our cohort [56] and not detected in any of the parents of probands 3 or 4. Proband 4 had a stage 4S neuroblastoma, which subsequently underwent spontaneous resolution, as was commonly observed for this type of tumour.[44] At that time, a predisposition for neoplasm development in WS and other overgrowth syndromes was already recognized.[44] This patient also presented with congenital heart defects. Recently, *de novo* mutations in histone-modifying genes have been implicated in many cases of congenital heart disease,[286] further supporting a role for *EZH2* in this individual's phenotype.

The clinical features of proband 5 are summarized in Table 2-3. The most striking and unusual feature is polymicrogyria. Photographs and MRI brain images are presented in Figure 2-2. We identified a c.2050C>T (p.Arg684Cys) mutation in *EZH2,* confirmed to be *de novo* by trio-based testing at the BCCH Molecular Genetics Lab. This mutation was previously described in five other WS patients,[58] and thus appears to be a relatively frequent cause of WS.

The clinical features of probands 6 and 7 are summarized in Table 2-3. In both unrelated patients we identified a c.398A>G (p.Tyr133Cys) *de novo* mutation in *EZH2*. This mutation was predicted damaging by PROVEAN/SIFT.

The clinical features of proband 8 are summarized in Table 2-3. We identified a c.403G>A (p.Gly135Arg) *de novo* mutation in *EZH2*. This mutation was predicted damaging by PROVEAN/SIFT. This patient also carried the c.553G>C (p.Asp185His) common variant in *EZH2*, inherited from his unaffected mother.

The clinical features of proband 9 are summarized in Table 2-3. We identified a c.1991G>A (p.Ser669Asn) mutation in *EZH2*. This mutation was predicted damaging by PROVEAN/SIFT. Interestingly, this mutation was inherited from the father, who also shows clinical features of WS; this represents the only confirmed familial case of molecularly diagnosed WS in our cohort.

The clinical features of proband 10 are summarized in Table 2-3. In blood-derived DNA, we identified three different variants in *EZH2*: the c.553G>C (p.Asp185His) common variant, the c.2233G>A, (p.Glu745Lys) rare variant previously described by Tatton-Brown *et al*. in a patient

with WS and lymphoma,[57] and a novel c.1323A>G (p.Glu441=) synonymous variant that appeared to be at less than 50% frequency by Sanger peak height, suggesting that it was somatic. As expected, p.(Glu745Lys) was predicted damaging by PROVEAN/SIFT while p.(Glu441=) was predicted to be neutral/tolerated. Because this patient had already developed leukemia at the time that the DNA was collected, we requested a DNA sample from a different tissue (excluding saliva which is also rich in lymphocytes). The medical team was able to collect nail clippings from the patient, and we successfully extracted DNA from these. Both p.(Asp185His) and p.(Glu745Lys) were present in the nail-derived DNA, whereas p.(Glu441=) was not, confirming the somatic nature of the latter variant (Figure 2-3); the percentage of mosaicism cannot be accurately estimated from the Sanger traces but appears to be around 15-20% based on peak height. Although the p.(Glu441=) variant is likely benign, this finding suggests that our assay is capable of picking up some level of somatic mosaicism. Furthermore, parental samples were also tested. As expected, p.(Glu441=) was not present in either parent. The common variant p.(Asp185His) was present in the unaffected mother while p.(Glu745Lys) was found to be a *de novo* mutation.

Finally, the clinical features of proband 11 are also summarized in Table 2-3. We validated the c.420T>A (p.Asp140Glu) rare variant in *EZH2* which had been detected externally by an overgrowth panel (from GeneDx). This variant was predicted damaging by PROVEAN but tolerated by SIFT. Unfortunately, parental samples were not available and thus inheritance could not be established; this variant remains classified as likely pathogenic until testing of parental samples is carried out.

| Characteristics | Proband 4 | Proband 5 | Proband 6 | Proband 7 | Proband 8 | Proband 9 | Proband 10 | Proband 11 |
|---|---|---|---|---|---|---|---|---|
| Sex | male | male | female | male | male | male | male | female |
| *EZH2* variant | c.394C>T, p.(Pro132Ser) | c.2050C>T, p.(Arg684Cys) | c.398A>G, p.(Tyr133Cys) | c.398A>G, p.(Tyr133Cys) | c.403G>A, p.(Gly135Arg) | c.1991G>A, p.(Ser669Asn) | c.2233G>A, p.(Glu745Lys) | c.420T>A, p.(Asp140Glu) |
| Inheritance | *de novo* | *de novo* | *de novo* | *de novo* | *de novo* | paternal (who is affected) | *de novo* | NK |
| **Growth features** | | | | | | | | |
| Gestational age at delivery (weeks) | 38 | 34 (preterm) | 39 | 40 + 3 days | 39-40 | 36 + 4 days | NK (term) | NK |
| Birth weight (kg) | 3.59 ($69^{th}$-$83^{rd}$ %ile) | 3.08 ($98^{th}$ %ile) | 3.53 ($72^{nd}$-$99^{th}$ %ile) | 4.35 ($92^{nd}$-$99^{th}$ %ile) | 3.99 ($81^{st}$-$99^{th}$ %ile) | 4.92 (>$99^{th}$ %ile) | 4.78 ($92^{nd}$-$99^{th}$ %ile) | NK |
| Birth length (cm) | 54.5 ($99^{th}$ %ile) | 49.5 ($97^{th}$ %ile) | 52.4 ($89^{th}$-$96^{th}$ %ile) | 57.5 (>$99^{th}$ %ile) | 55 ($96^{th}$-$99^{th}$ %ile) | 53.5 (>$99^{th}$ %ile) | 58 (>$99^{th}$ %ile) | NK |
| Birth head circumference (cm) | 37 ($98^{th}$ %ile) | 33 ($90^{th}$ %ile) | 36.2 ($93^{rd}$-$97^{th}$ %ile) | 37 ($89^{th}$-$98^{th}$ %ile) | 36 ($75^{th}$-$89^{th}$ %ile) | 37 (>$99^{th}$ %ile) | 38 ($92^{nd}$-$98^{th}$ %ile) | NK |
| Recent weight (kg) [age measured] | 16 [13m] (>$99^{th}$ %ile) | 22 [2y7m] (>$99^{th}$ %ile) | 69 [17y] ($87^{th}$ %ile) | 22 [2y6m] (>$99^{th}$ %ile) | 64 [10y10] (>$99^{th}$ %ile) | 17.3 [14m] (>$99^{th}$ %ile) | 21 [2y4m] (>$99^{th}$ %ile) | 17.9 [3y11m] ($83^{rd}$ %ile) |
| Recent height (cm) [age measured] | 91.4 [13m] (>$99^{th}$ %ile) | 102.7 [2y7m] (>$99^{th}$ %ile) | 185 [17y] (>$99^{th}$ %ile) | 105 [2y6m] (>$99^{th}$ %ile) | 162.3 [10y10] (>$99^{th}$ %ile) | 92.8 [14m] (>$99^{th}$ %ile) | 107 [2y4m] (>$99^{th}$ %ile) | 99 [3y11m] ($39^{th}$ %ile) |
| Recent head circumference (cm) [age measured] | 51 [13m] (>$99^{th}$ %ile) | 49.4 [2y7m] ($51^{st}$ %ile) | NK | NK | NK | 50 [14m] (>$99^{th}$ %ile) | 53 [2y4m] (>$99^{th}$ %ile) | 59.5 [3y11m] (>$99^{th}$ %ile) |
| Excessive growth of prenatal onset | +++ | +++ | +++ | +++ | +++ | +++ | +++ | + (head only) |
| Tall stature | +++ | +++ | +++ | +++ | +++ | +++ | ++ | - |
| Accelerated osseous maturation | ++ | + (18m at 10y6m) | ++ (3y6m at 1y8m) | NK | + | NK | +++ (6y at 2y4m) | NK |
| **Neurological features** | | | | | | | | |
| Hypertonia | + (knees) | + (peripheral) | - | + | + | - | - | - |
| Hypotonia | NK | + (abdominal, left side more prominent) | + | - | ++ | + | + | - |
| Hoarse low-pitched cry | ++ | - | ++ | - | - | NK | + | NK |
| Intellectual disability | + | + | + | + | + | NK | NK | NK |
| Developmental | - | +++ | + | - | + | ++ | + | + |

| Characteristics | Proband 4 | Proband 5 | Proband 6 | Proband 7 | Proband 8 | Proband 9 | Proband 10 | Proband 11 |
|---|---|---|---|---|---|---|---|---|
| delay | | | | | | | | |
| Speech delay | + (mild) | +++ | + | + | + | + | ++ | ++ |
| Behavioural problems | NK | - | - | + | - | NK | NK | - |
| Excessive appetite | NK | + | - | ++ | - | NK | NK | - |
| Ventriculomegaly | NK | - | - | - | - | - | NK | - |
| Delayed myelination | NK | - | - | NK | NK | - | NK | - |
| Cerebellar hypoplasia | NK | - | - | - | - | - | NK | - |
| Seizures [onset] | NK | 1 GTC febrile-associated [~ 9m] | - | - | ++ epilepsy [3y] | - | NK | - |
| Polymicrogyria | - | +++ asymmetric perisylvian | - | - | NK | - | NK | - |
| Pachygyria | - | - | - | - | ++ | - | NK | - |
| Poor fine motor coordination | + | + | + | - | + | + | NK | + |
| Poor balance/gravitational insecurity | ++ | + | ++ | - | + | NK | NK | - |
| **Craniofacial** | | | | | | | | |
| Macrocephaly | +++ | - | +++ (> +2 S.D. at 2y1m) | - (+1.5 S.D. at 7m) | ++ | +++ | + | ++ |
| Large bifrontal diameter | ++ | - | ++ | + | + | + | + | + |
| Flat occiput | ++ | + | + | + | + | + | NK | NK |
| Large ears | ++ (with hearing loss) | +++ | ++ | + | + | ++ | - | - |
| Ocular hypertelorism | ++ | ++ | ++ | ++ | + | + | + | + |
| Down slanted palpebral fissures | + | + | + | + | + | - | - | - |
| Long philtrum | ++ | - | + | - | + | + | + | - |
| Micro/retrognathia | + | ++ | + | + | - | + | + | + |

| Characteristics | Proband 4 | Proband 5 | Proband 6 | Proband 7 | Proband 8 | Proband 9 | Proband 10 | Proband 11 |
|---|---|---|---|---|---|---|---|---|
| **Cardiovascular** | | | | | | | | |
| Patent ductus arteriosus | ++ | - | - | - | NK | - | NK | NK |
| Ventricular septal defect | ++ | - | - | - | NK | - | NK | NK |
| **Limbs** | | | | | | | | |
| Limited elbow and knee extension in early life | + | + | - | - | + | - | - | NK |
| Limited elbow and knee extension after puberty | NK | n/a | - | n/a | n/a | n/a | n/a | n/a |
| Widened distal femurs and ulnas | ++ | - | NK | NK | NK | NK | NK | NK |
| **Hands** | | | | | | | | |
| Prominent digit pads | + | - | + | + | + | - | - | NK |
| Single transverse palmar crease | ++ | ++ | - | + | - | - | - | NK |
| Camptodactyly | NK | - | + | - | - | - | - | NK |
| Broad thumbs | NK | - | - | - | + | - | - | NK |
| Thin, deep-set nails | - | ++ | + | - | - | + | - | NK |
| **Feet** | | | | | | | | |
| Clinodactyly, toes | + | - | - | - | - | - | NK | - |
| Talipes equinovarus | + | - | - | - | + | - | NK | NK |
| Short fourth metatarsals | NK | - | - | - | - | NK | NK | NK |
| Hind foot valgus | NK | - | - | - | - | - | NK | NK |
| **Skin** | | | | | | | | |
| Excessive loose skin | ++ | + | + | - | - | + | - | + |
| Hypoplastic/ supernumerary nipples | + | + | - | - | - | - | + | + |
| Thin hair | - | + | - | - | - | + | NK | - |
| Increased pigmented nevi | + | - | - | - | - | + (3 café-au-lait spots) | NK | - |
| **Connective tissue** | | | | | | | | |
| Umbilical hernia | + | - | + | + | ++ | + | + | NK |
| Inguinal hernia | NK | - | - | - | - | - | - | NK |

| Characteristics | Proband 4 | Proband 5 | Proband 6 | Proband 7 | Proband 8 | Proband 9 | Proband 10 | Proband 11 |
|---|---|---|---|---|---|---|---|---|
| Diastasis recti | + | - | - | + | - | + | + | NK |
| Scoliosis | NK | mild, resolved | - | - | + (NK) | - | NK | - |
| Kyphosis | + | - | + | - | - | - | NK | - |
| **Endocrine** | | | | | | | | |
| Hypothyroidism [onset] | NK | - | - | - | - | - | NK | NK |
| Growth hormone deficiency [onset] | NK | - | - | - | - | NK | NK | NK |
| Hypoglycemia [onset] | perinatal | mild, resolved [birth] | - | - | - | + (resolved) [neonatal] | NK | NK |
| **Neoplasia** | | | | | | | | |
| Neuroblastoma [onset] | regressed [birth] | - | removed [prenatal] | - | - | - | NK | - |
| Leukemia [onset] | - | - | - | - | - | - | + (ALL) [11m] | - |
| Lymphoma [onset] | - | - | - | - | - | - | - | - |

**Table 2-3: Phenotypic manifestations of Weaver syndrome in patients with *EZH2* mutations.**

+ = minimally present; ++ = obviously present; +++ = very prominent; - = assessed and found to be absent; NK = not known; n/a = non applicable; y = years; m = months; %ile = percentile; SD = standard deviation(s); GTC = generalized tonic-clonic; ALL = acute lymphoblastic leukemia.


**Figure 2-2: Weaver syndrome proband with polymicrogyria described in this study.**

Proband 5 is shown at 2 months, 4 months, 6 months, 8 months, 12 months and 19 months (a). Both sides of the hand are shown at 12 months (b) to illustrate the prominent palmar crease. At 27 months (c), face with prominent rosy cheeks, profile and ears are shown to confirm the dimple is only present behind the right ear. At 31 months, mild camptodactyly is seen on the toes and a third nipple is apparent (d). Full torso and full body are also shown at 31 months (e). X-rays of the hand at 11½ months and the knee at 10½ months (f) are indicative of advanced bone age. MRI done at 5 days of age (g) illustrates asymmetric perisylvian polymicrogyria.

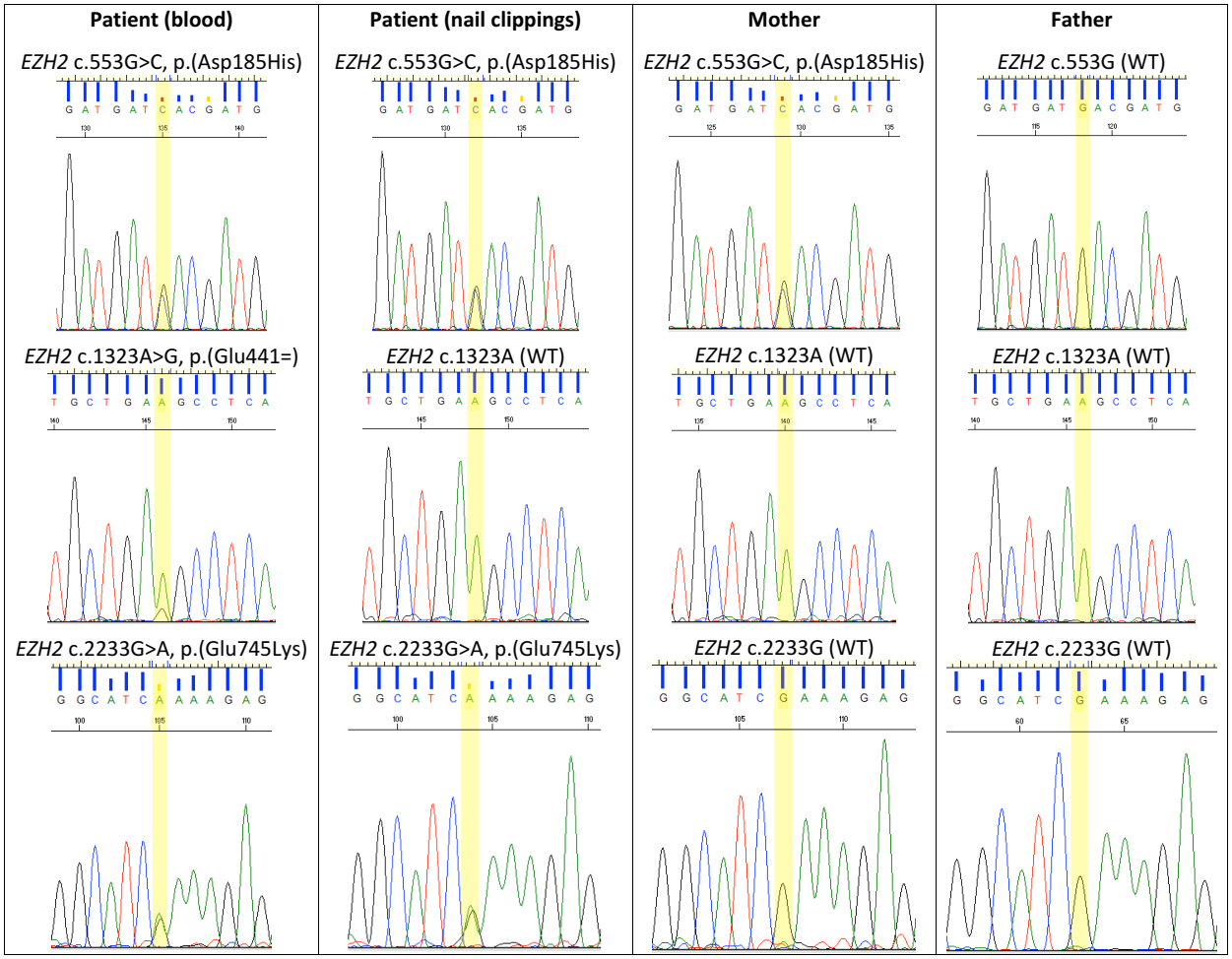



**Figure 2-3: Sanger *EZH2* results and validations for proband 10.**

Sanger sequencing identified three different *EZH2* variants in proband 10. Because the proband had developed leukemia, we validated the variants in a tissue other than blood (nail clippings) and found that p.(Glu441=) is a somatic variant, possibly cancer-specific. Further, we validated all the variants in parental samples and found that the common variant p.(Asp185His) was present in the unaffected mother while the rare variant p.(Glu745Lys) was not present in either parent, confirming its sporadic nature and supporting pathogenicity. Sanger traces were analyzed using Sequence Scanner v1.0. WT = wild-type.

#### 2.3.2.2.2 Common variant in *EZH2* detected in our overgrowth cohort

In addition to the variants described above, we also detected the c.553G>C (p.(Asp185His) or D185H) variant in *EZH2* in an additional six individuals referred to our WS-like cohort (cases 15, 40, 53, 73, 118, 119). Note that cases 118 and 119 are two affected siblings. This variant was

also found in cases 95 and 130 via external testing, and validated independently in our laboratory. Clinical features of these eight individuals are summarized in Table 2-4. Other members of their families were tested at this locus and in all cases the variant was found to be inherited from one of the parents (Table 2-4). Since this *EZH2* variant was also present in probands 8 and 10, in conjunction with a pathogenic mutation in the same gene, the total frequency of the D185H variant within our patient cohort adds up to 15% (10/66).

When we first detected D185H in cohort members who presented with overgrowth and dysmorphic features, minimal population frequency information was available thus warranting further investigation prior to classification. Currently, D185H is reported in the dbSNP database (rs2302427C>G) with a minor allele frequency of 8% in the 1000 Genomes population, and is reported in other databases with similar frequencies (see Table 2-2): 6% in the Exome Variant Server, 7.9% in ExAc, and 7.7% in a healthy ancestrally diverse cohort screened for common variants in cancer-susceptibility genes.[287] This variant lies within a stretch of seven aspartic acid residues in a row, ranging from residue 183 to 189. In addition to the D185H common variant, rare variants have been described in dbSNP affecting nearly all of these residues (see Appendix B), suggesting a certain tolerance for alterations within this amino acid repeat sequence.

Because of its frequency in the general population, the D185H variant cannot in isolation be causative of WS. Although this variant appears at a frequency of 15% in our cohort, such apparent enrichment is likely due to chance, or possibly population stratification. We know that at least 62% of our patients are Caucasians, which are well represented in the 1000 Genomes and ExAc projects, and closely match the European American population from the Exome Variant Server in whom this variant is reported at a frequency of 8.4%. However, when looking more closely, six of the ten D185H carriers are of European ancestry, reflecting a 9.1% frequency in this population, similar to that reported. Among the other four carriers, two are from South America and the other two are likely to be of "latino" descent; these populations are not as well represented in the "big data" projects.

Interestingly, D185H is predicted neutral by PROVEAN but damaging by SIFT. Based solely on population frequency, this variant must be "benign" relative to WS, though it may not be benign with reference to all possible disease phenotypes. To resolve this apparent discrepancy, we decided to include the D185H variant in our functional work described in Chapter 3.

| **Characteristics** | **Case 15** | **Case 40** | **Case 53** | **Case 73** | **Case 95** | **Case 118** | **Case 119** | **Case 130*** |
|---|---|---|---|---|---|---|---|---|
| Sex | male | female | female | male | male | female | male | female |
| *EZH2* variant | c.553G>C; p.(Asp185His) | | | | | | | |
| Inheritance | paternal (unaffected) | paternal (unaffected) | maternal (unaffected) | maternal (unaffected) | maternal (unaffected) | paternal (unaffected) | paternal (unaffected) | paternal (unaffected) |
| Other carriers in the family | sibling (unaffected); paternal aunt (unaffected) | n/a | sibling (unaffected) | n/a | n/a | sibling (case 119) | sibling (case 118) | n/a; note that affected sibling is not a carrier |
| **Growth features** | | | | | | | | |
| Gestational age at delivery (weeks) | 38 | 38 | 38 | 38 | 39 | NK | NK | NK (term) |
| Birth weight (kg) | 3.45 ($58^{th}$-$74^{th}$ %ile) | 3.55 ($75^{th}$-$84^{th}$ %ile) | 4.45 ($99^{th}$ %ile) | 3.05 ($34^{th}$-$50^{th}$ %ile) | 3.8 ($82^{nd}$ %ile) | NK | NK | 3.99 ($64^{th}$-$86^{th}$ %ile) |
| Birth length (cm) | 50 ($52^{nd}$-$62^{nd}$ %ile) | 52 ($92^{nd}$-$94^{th}$ %ile) | 49 ($47^{th}$-$57^{th}$ %ile) | 52 ($92^{nd}$-$94^{th}$ %ile) | 50 ($46^{th}$-$52^{nd}$ %ile) | NK | NK | NK (>$75^{th}$ %ile) |
| Birth head circumference (cm) | 34 ($36^{th}$-$52^{nd}$ %ile) | NK | 35 ($84^{th}$ %ile) | 35 ($84^{th}$ %ile) | 36 ($85^{th}$-$89^{th}$ %ile) | NK | NK | NK ($99^{th}$ %ile) |
| Recent weight (kg) [age measured] | 78 [15y5m] ($93^{rd}$ %ile) | 55 [10y2m] ($98^{th}$ %ile) | 45 [9y6m] ($95^{th}$ %ile) | 85.6 [9y11m] (>$99^{th}$ %ile) | 31.7 [5y6m] (>$99^{th}$ %ile) | NK | 15.1 [3y3m] ($58^{th}$ %ile) | 42.8 [9y3m] ($95^{th}$ %ile) |
| Recent height (cm) [age measured] | 180 [15y5m] ($87^{th}$ %ile) | 154 [10y2m] ($98^{th}$ %ile) | 150 [9y6m] ($98^{th}$ %ile) | 155 [9y11m] (>$99^{th}$ %ile) | 122.8[5y6m] (>$99^{th}$ %ile) | NK | 93 [3y3m] ($16^{th}$ %ile) | 141.8 [9y3m] ($88^{th}$ %ile) |
| Recent head circumference (cm) [age measured] | 56.5 [15y5m] (~ $85^{th}$ %ile) | NK | NK | 56.2 [9y11m] ($99^{th}$ %ile) | NK | NK | 54 [3y3m] (>$99^{th}$ %ile) | NK [9y3m] (>+3SD) |
| Excessive growth of prenatal onset | - | + | + | + | + | NK | + | ++ |
| Tall stature | ++ | +++ | +++ | +++ | +++ | NK | - | + |
| Accelerated osseous maturation | ++ (14-15y at 12y2m) | NK | ++ (5y9m at 4y) | NK | + (3y6m at 2y8m) | NK | - (10-13m at 10m) | NK |
| **Neurological features** | | | | | | | | |
| Hypertonia | - | +/- | - | - | - | - | - | - |
| Hypotonia | - | +/- | + | + | + | +++ | +++ | + |
| Hoarse low-pitched cry | - | ++ | + | - | NK | NK | NK | NK |
| Intellectual disability | + | + | ++ | + | + | NK | NK | + |

| Characteristics | Case 15 | Case 40 | Case 53 | Case 73 | Case 95 | Case 118 | Case 119 | Case 130* |
|---|---|---|---|---|---|---|---|---|
| Developmental delay | + | + | + | + | + | + | + | ++ |
| Speech delay | + | NK | + | - | ++ | + | + | +++ |
| Behavioural problems | + | + | NK | - | + | NK | + | ++ |
| Autism spectrum disorder | - | NK | NK | ++ | - | +++ | ++ | ++ |
| Excessive appetite | - | NK | ++ | ++ | + | NK | NK | - |
| Ventriculomegaly | - | + | NK | - | - | NK | NK | - |
| Delayed myelination | - | - | NK | - | - | NK | NK | ++ |
| Cerebellar hypoplasia | - | - | NK | - | - | NK | NK | - |
| Seizures [onset] | - | - | - | - | - | NK | NK | - |
| Polymicrogyria | - | - | NK | - | - | NK | NK | - |
| Pachygyria | - | - | NK | - | - | NK | NK | - |
| Other brain abnormalities | - | cystic lesion in the area of the lateral ventricle | NK | thinning of the inner parietal bones | - | peri-ventricular leuko-malacia | mild external hydro-cephalus | mild to moderate megalen-cephaly (+3.3 S.D.) |
| Poor fine motor coordination | +++ | ++ | + | - | + | NK | NK | + |
| Poor balance/gravitational insecurity | ++ | - | - | - | + | NK | NK | - |
| **Craniofacial** | | | | | | | | |
| Macrocephaly | ++ | - | - | +++ | ++ | NK | +++ | +++ |
| Large bifrontal diameter | + | - | + | - | - | NK | NK | NK |
| Flat occiput | ++ | - | - | - | - | NK | - | NK |
| Large ears | - | - | + | - | - | NK | - | - |
| Ocular hypertelorism | ++ | - | ++ | - | + | NK | + | - |
| Down slanted palpebral fissures | - | - | + | - | + | NK | NK | - |
| Long philtrum | - | - | + | - | - | NK | + | - |
| Micro/retrognathia | - | + | - | - | + | NK | - | - |

| Characteristics | Case 15 | Case 40 | Case 53 | Case 73 | Case 95 | Case 118 | Case 119 | Case 130* |
|---|---|---|---|---|---|---|---|---|
| **Cardiovascular** | | | | | | | | |
| Patent ductus arteriosus | - | NK | - | NK | NK | NK | NK | NK |
| Ventricular septal defect | - | NK | - | NK | NK | NK | NK | NK |
| **Limbs** | | | | | | | | |
| Limited elbow and knee extension in early life | + | NK | - | - | NK | NK | - | - |
| Limited elbow and knee extension after puberty | + | NK | n/a | n/a | n/a | n/a | n/a | n/a |
| Widened distal femurs and ulnas | NK | - | - | NK | NK | NK | NK | NK |
| **Hands** | | | | | | | | |
| Prominent digit pads | - | - | + | - | - | NK | NK | NK |
| Single transverse palmar crease | - | NK | - | - | - | NK | NK | + (left) |
| Camptodactyly | - | - | + | - | - | NK | NK | - |
| Broad thumbs | + | - | + | - | - | NK | NK | NK |
| Thin, deep-set nails | - | - | + | - | - | NK | NK | - |
| **Feet** | | | | | | | | |
| Clinodactyly, toes | - | + | + | - | - | NK | NK | - |
| Talipes equinovarus | - | + | - | - | - | NK | NK | - |
| Short fourth metatarsals | + | - | - | - | - | NK | NK | - |
| Hind foot valgus | + | + | - | - | - | NK | NK | - |
| **Skin** | | | | | | | | |
| Excessive loose skin | - | - | - | - | - | NK | NK | - |
| Hypoplastic/ supernumerary nipples | - | NK | - | - | NK | NK | NK | - |
| Thin hair | - | - | - | - | NK | NK | NK | NK |
| Increased pigmented nevi | - | NK | - | linear skin hyper-pigmentation across back | - | NK | NK | - |
| **Connective tissue** | | | | | | | | |
| Umbilical hernia | + | NK | ++ | - | NK | NK | NK | - |

| Characteristics | Case 15 | Case 40 | Case 53 | Case 73 | Case 95 | Case 118 | Case 119 | Case 130* |
|---|---|---|---|---|---|---|---|---|
| Inguinal hernia | - | NK | - | - | NK | NK | NK | - |
| Diastasis recti | - | NK | - | - | NK | NK | NK | - |
| Scoliosis | mild | NK | - | - | NK | NK | NK | - |
| Kyphosis | - | - | - | - | - | NK | NK | - |
| **Endocrine** | | | | | | | | |
| Hypothyroidism [onset] | - | - | - | - | NK | NK | NK | NK |
| Growth hormone deficiency [onset] | - | - | - | - | NK | NK | NK | NK |
| Hypoglycemia [onset] | - | - | - | - | NK | NK | NK | NK |
| **Neoplasia** | | | | | | | | |
| Neuroblastoma [onset] | - | - | NK | - | - | NK | NK | - |
| Leukemia [onset] | - | - | - | - | - | NK | NK | - |
| Lymphoma [onset] | - | - | - | - | - | NK | NK | - |

**Table 2-4: Phenotypic description of carriers for the p.(Asp185His) polymorphism identified within our cohort.**

***** Case 130 was referred to our cohort (with affected sibling) because she was a carrier of this *EZH2* variant, and not because of the clinical diagnosis, which was not suggestive of Weaver syndrome. Data on this patient is included here for completeness.

+ = minimally present; ++ = obviously present; +++ = very prominent; - = assessed and found to be absent; %ile = percentile; S.D. = standard deviation(s); +/- = tone reported as abnormal, but hypertonia or hypotonia not specified; NK = not known; n/a = non applicable; y = years; m = months.

#### 2.3.2.3 Mutations and variants identified in *NSD1*

With 45 individuals in whom we did not find any pathogenic variants or VOUS in *EZH2*, and two individuals who were not tested for *EZH2* variants because their phenotype was more consistent with Sotos syndrome, 47 patients remained to diagnose. By this time, two individuals had been found externally to have an alternative diagnosis, such that no further investigations were required on our part. This left us with 45 individuals to consider for *NSD1* testing. Of these 45, 11 had been previously been tested for variants in *NSD1*; six had been found to be negative (three via single gene testing and three via exome sequencing), and the remaining five were reported to have either pathogenic variants (2/5) or VOUS (3/5) in *NSD1*, which I validated independently. The entire coding region of *NSD1* was sequenced in 20 out of the remaining 34 individuals. The other 14 patients were not tested because their phenotype was not consistent with Sotos syndrome and thus testing of this larger gene was not considered to be cost-effective.

All variants identified or validated in the coding region of *NSD1*, or within 20 nucleotides either 5' or 3' of the exonic sequences, are summarized in Table 2-5. Other common variants located further away from the intron/exon boundaries that were observed upon Sanger analysis were included in the individual reports for completeness but are not included in Table 2-5. Pathogenic mutations and likely pathogenic variants identified in *NSD1* are scattered throughout the gene, as illustrated in Figure 2-4. For description purposes, these individuals were renamed as probands 12 to 18 and their clinical features are summarized in Table 2-6. Patients with confirmed pathogenic mutations in *NSD1* were considered diagnosed and not investigated further.

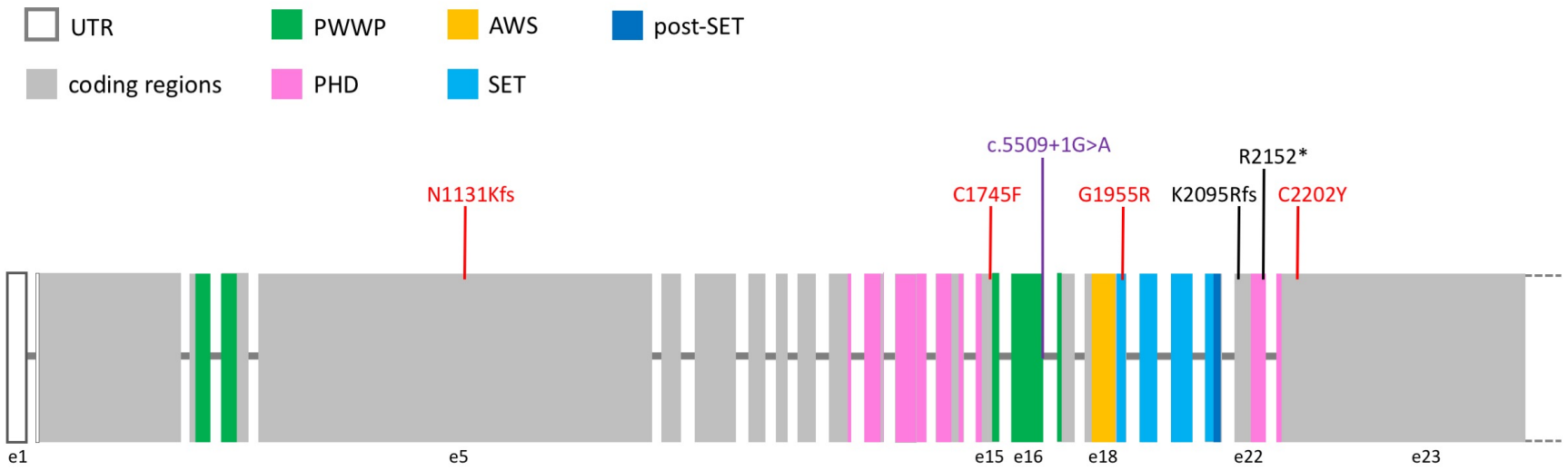



**Figure 2-4: Schematic representation of human NSD1.**

Human NSD1 is represented. Each rectangle represents one exon. Exon size is represented to scale, while intronic distances are not to scale. Exon 23 is very large thus only the coding region (approximately ¼ of the exon) is represented here. White (open) rectangles represent non-coding UTRs and grey rectangles represent coding exons (NM_022455.4). NSD1 protein contains 2696 amino acids (NP_071900.2) and ten recognizable domains (two PWWP, five PHD, and one each of AWS, SET and post-SET domains), represented here in coloured rectangles according to NCBI (NP_071900.2) and UniProtKB/InterProt (Q96L73-1) coordinates. Variants in black represent pathogenic mutations previously known and validated in our laboratory; variants in red represent newly identified pathogenic mutations; variant in purple represents a likely pathogenic splice site variant. e = exon; UTR = untranslated region; PWWP = proline- tryptophan-tryptophan-proline; PHD = (plant homeodomain) zinc finger motif; AWS = associated with SET domains; SET = *Su(var)3-9, E(z) and Trithorax* domain.

| Genomic position (chr 5, GRCh37.p13) | State | Position within gene structure | Predicted protein change | Presence in public databases and/or corresponding MAFs when applicable | | | | | | | Interpretation | Prevalence among tested |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | dbSNP [a] | EVS (EA/All) | ExAc | Clin Var ID | DECIPHER | NSD1 LOVD (v2.0) | COSMIC (COSM number) | | |
| 176636882 | het<br>homo | exon 5 (c.1482C>T) | C494= | rs1363405 (46.8%) | 19/ 35.4% | 30% | 96035 | nds | 00069 | nds | benign | 2/20<br>1/18 |
| 176637149 | het | exon 5 (c.1749G>A) | E583= | rs3733874 | 11.3/ 10% | 17.7% | 96038 | nds | 00039 | nds | benign | 1/20 |
| 176637240 | het | exon 5 (c.1840G>T) | V614L | rs3733875 (22.9%) | 11.4/ 9.9 % | 17.6% | 96042 | nds | 00041 | 4416947 | benign | 1/20 |
| 176637576 | homo | exon 5 (c.2176T>C) | S726P | rs28932178 (25%) | 13.8/ 12.6% | 19.9% | 96043 | nds | 00043 | nds | benign | 1/20 |
| 176638489 and 176638490 | het<br>het | exon 5 (c.3089T>C) and exon 5 (c.3090G>T) in phase | L1030S and L1030= | rs200856103 and rs201860097 | nds; 0.01/ <0.01% | <0.01% each, in phase in 52 people | 159299; nds | nds; nds | nds; nds | nds; nds | likely benign (both) | 2/22 |
| 176638793 | het | exon 5 (c.3393delC) | N1131Kfs | nds | nds | nds | nds | nds | nds | nds | likely pathogenic [b] | 1/20 |
| 176639105 | het | exon 5 (c.3705T>C) | N1235= | rs28932181 (10.6%) | 5/11.4% | 7% | 96057 | nds | 00046 | nds | benign | 2/20 |
| 176694650 | het | exon 15 (c.5234G>T) | C1745F | nds | nds | nds | nds | nds | nds | nds | pathogenic | 1/20 |
| 176696809 | het | flanking exon 16 (c.5509+1G>A) | n/a | nds | nds | nds | nds | nds | nds | nds | likely pathogenic | 1/20 |
| 176707806 | het | exon 18 (c.5863G>C) | G1955R | nds | nds | nds | nds | nds | nds | nds | pathogenic | 1/20 |
| 176718980 | het | exon 22 (c.6284delA) [c] | K2095Rfs | nds | nds | nds | nds | nds | nds | nds | pathogenic | 1/20 |
| 176719150 | het | exon 22 (c.6454C>T) | R2152* | rs587784199 | nds | nds | 159421 | nds | nds | nds | pathogenic | 1/20 |
| 176720974 | het | exon 23 (c.6605G>A) | C2202Y | rs121908071 | nds | nds | 4146 | nds | 000105 (v3.0) | nds | pathogenic | 1/20 |
| 176721119 | het | exon 23 (c.6750G>A) | M2250I | rs35848863 (5.2%) | 4.9/5.2% | 4.9% | 96070 | nds | 00050 | nds | benign | 1/20 |
| 176721151 | het | exon 23 (c.6782T>C) | M2261T | rs34165241 (5.2%) | 4.9/5.2% | 4.9% | 96071 | nds | 00051 | nds | benign | 1/20 |
| 176721198 | homo<br>het | exon 23 (c.6829T>C) | L2277= | rs28580074 (22.9%) | 11.5/ 10.1% | 82.4% | 96073 | nds | 00052 | nds | benign | 18/20<br>2/2 |
| 176721272 | het<br>homo | exon 23 (c.6903G>C) | G2301= | rs11740250 (10.8%) | 25.1/ 18% | 19.2% | 96074 | nds | 00053 | nds | benign | 5/20<br>1/15 |
| 176721529 | het | exon 23 (c.7160C>T) | P2387L | rs766700264 | nds | <0.01% | nds | nds | nds | nds | VOUS | 1/20 |
| 176722005 | het | exon 23 (c.7636G>A) | A2546T | rs78247455 (27.4%) | nds | 2.6% | 96079 | nds | 00074 | nds | benign | 1/20 |
| 176722219 | het | exon 23 (c.7850T>C) | L2617S | rs77618751 (<0.01%) | 0.37/ 0.28% | <0.01% | 96082 | nds | 00075 | nds | VOUS | 1/21 |

**Table 2-5: List of variants identified near or within the coding regions of *NSD1*.**

[a] Minor allele frequencies (MAFs) provided for dbSNP correspond to the values given for the 1000 Genomes project.

[b] Not yet validated in parental samples.

[c] This *NSD1* alteration was reported externally as (p.Ser2096Valfs*17).

Blue = exclusively previously identified variants that have been validated; het = heterozygous; homo = homozygous; n/a = non applicable; nds = not described; EA = European American population.

| **Characteristics** | **Proband 12** | **Proband 13** | **Proband 14** | **Proband 15** | **Proband 16** | **Proband 17** | **Proband 18** |
|---|---|---|---|---|---|---|---|
| Sex | female | female | male | male | male | male | female |
| *NSD1* variant | c.6284delA, p.(Lys2095Argfs) | c.6454C>T, p.(Arg2152*) | c.5234G>T, p.(Cys1745Phe) | c.5863G>C, p.(Gly1955Arg) | c.6605G>A, p.(Cys2202Tyr) | c.5509+1G>A | c.2586delC, p.(Asn1131Lysfs) |
| Inheritance | *de novo* | *de novo* | *de novo* | *de novo* | *de novo* | *de novo* | NK |
| **Growth features** | | | | | | | |
| Gestational age at delivery (weeks) | NK | 38.5 | NK | NK | NK | NK | 38 |
| Birth weight (kg) | NK ($75^{th}$ %ile) | 4.56 ($99^{th}$ %ile) | NK | NK | NK | NK | 3.47 |
| Birth length (cm) | NK ($98^{th}$ %ile) | 57.15 (>$99^{th}$ %ile) | NK | NK | NK | NK | NK |
| Birth head circumference (cm) | NK ($50^{th}$ %ile) | NK ("large") | NK | NK | NK | NK | NK |
| Recent weight (kg) [age measured] | NK [6y6m] ($90^{th}$-$97^{th}$ %ile) | 24.5 [3y4m] (>$99^{th}$ %ile) | 47.5 [11y] ($90^{th}$ %ile) | 19.1 [2y6] (>$99^{th}$ %ile) | NK | 62.4 [16y6m] ($48^{th}$ %ile) | 68.8 [27y7m] (~$80^{th}$ %ile) |
| Recent height (cm) [age measured] | NK [6y6m] ($50^{th}$ %tile) | 116.5 [3y4m] (>$99^{th}$ %ile) | 164.8 [11y] (>$99^{th}$ %ile) | 104.2 [2y6m] (>$99^{th}$ %ile) | NK | 186.8[16y6m] ($96^{th}$ %ile) | 176 [27y7m] (>$90^{th}$ %ile) |
| Recent head circumference (cm) [age measured] | NK [6y6m] ($95^{th}$ %tile) | NK (>$99^{th}$ %tile) | 58 [11y] (>$99^{th}$ %ile) | 52.8 [2y6m] ($99^{th}$ %ile) | NK | NK | 58.4 [27y7m] (>$95^{th}$ %ile) |
| Excessive growth of prenatal onset | ++ | +++ | NK | NK | NK | NK | NK |
| Tall stature | - | +++ | +++ | +++ | NK | +++ | ++ |
| Accelerated osseous maturation | + | +++ | NK | ++ | NK | NK | + |
| **Neurological features** | | | | | | | |
| Hypertonia | NK | - | - | - | NK | NK | - |
| Hypotonia | NK | - | + | - | NK | NK | + |
| Hoarse low-pitched cry | NK | NK | NK | + | NK | NK | NK |
| Intellectual disability | NK | - | NK | NK | NK | + | - |
| Developmental delay | + | - | NK | + | NK | + | + |
| Speech delay | + | + | NK | ++ | NK | NK | ++ |
| Behavioural problems | + | + | NK | + | NK | NK | ++ |
| Excessive appetite | NK | - | - | NK | NK | - | NK |
| Brain abnormalities | NK | - | NK | NK | NK | NK | NK |
| Seizures [onset] | - | - | NK | NK | NK | NK | - |
| Poor fine motor | NK | - | NK | NK | NK | + | + |

| Characteristics | Proband 12 | Proband 13 | Proband 14 | Proband 15 | Proband 16 | Proband 17 | Proband 18 |
|---|---|---|---|---|---|---|---|
| coordination | | | | | | | |
| Poor balance/ gravitational insecurity | NK | - | NK | NK | NK | NK | NK |
| **Craniofacial** | | | | | | | |
| Macrocephaly | +++ | +++ | +++ | +++ | + | ++ | +++ |
| Large bifrontal diameter | - | NK | - | + | + | - | - |
| Flat occiput | NK | NK | - | NK | + | - | NK |
| Large ears | + | - | + | + | - | + | + |
| Ocular hypertelorism | + | - | + | + | + | ++ | NK |
| Down slanted palpebral fissures | + | + | + | + | + | + | + |
| Long philtrum | - | NK | + | + | - | - | NK |
| Micro/retrognathia | + | NK | - | + | + | - | NK |
| **Cardiovascular** | | | | | | | |
| Patent ductus arteriosus | NK | - | NK | - | NK | NK | NK |
| Ventricular septal defect | NK | - | NK | - (dyskinesia) | NK | NK | NK |
| **Limbs** | | | | | | | |
| Limited elbow and knee extension in early life | - | - | - | - | NK | - | NK |
| Limited elbow and knee extension after puberty | n/a | n/a | n/a | n/a | NK | NK | - |
| Widened distal femurs and ulnas | NK | NK | NK | NK | NK | NK | NK |
| **Hands** | | | | | | | |
| Prominent digit pads | + | NK | - | NK | NK | NK | NK |
| Single transverse palmar crease | - | NK | - | NK | NK | NK | - |
| Camptodactyly | - | NK | - | - | NK | - | - |
| Broad thumbs | - | NK | - | NK | NK | - | NK |
| Thin, deep-set nails | NK | NK | - | NK | NK | NK | NK |
| **Feet** | | | | | | | |
| Clinodactyly, toes | - | - | - | - | NK | - | - |
| Talipes equinovarus | NK | - | - | NK | NK | NK | - |

| Characteristics | Proband 12 | Proband 13 | Proband 14 | Proband 15 | Proband 16 | Proband 17 | Proband 18 |
|---|---|---|---|---|---|---|---|
| Short fourth metatarsals | NK | - | - | NK | NK | NK | - |
| Hind foot valgus | NK | - | - | NK | NK | + | - |
| **Skin** | | | | | | | |
| Excessive loose skin | - | - | - | + | NK | - | - |
| Hypoplastic/ supernumerary nipples | NK | NK | + | NK | + | NK | NK |
| Thin hair | - | NK | NK | NK | NK | - | + |
| Increased pigmented nevi | NK | - | + (4 café-au-lait spots) | - | NK | NK | - |
| **Connective tissue** | | | | | | | |
| Umbilical hernia | NK | NK | + | NK | NK | NK | NK |
| Inguinal hernia | NK | NK | + | NK | NK | NK | NK |
| Diastasis recti | NK | - | - | NK | NK | - | NK |
| Scoliosis | - | - | - | NK | NK | ++ | + |
| Kyphosis | NK | - | - | - | NK | - | - |
| **Endocrine** | | | | | | | |
| Hypothyroidism [onset] | - | - | NK | NK | NK | - | NK |
| Growth hormone deficiency [onset] | - | - | NK | NK | NK | NK | NK |
| Hypoglycemia [onset] | - | + [perinatal, resolved] | NK | NK | NK | NK | NK |
| **Neoplasia** | | | | | | | |
| Type [age of onset] | NK | neuroblastoma [between 2y6m and 3y] | NK | NK | NK | - | - |

**Table 2-6: Phenotypic manifestations of Sotos syndrome in patients with *NSD1* mutations.**

+ = minimally present; ++ = obviously present; +++ = very prominent; - = assessed and found to be absent; NK = not known; n/a = non applicable; y = years; m = months; %ile = percentile**.**

## 2.4 Conclusions based on detailed phenotyping and targeted gene sequencing approach

### 2.4.1 Clarifying the Weaver phenotype

The amount of information available varied in quantity and quality between referrals, and the classification of many traits (that do not involve quantitative measurements) was subjective and dependent on the familiarity of the physician with rare overgrowth and dysmorphic features. Nonetheless, the information collected was still useful for our analysis.

Table 2-7 shows the prevalence of phenotypic traits within different sub-groups of our overgrowth cohort: *EZH2* "positive", *NSD1* "positive", and *EZH2*/*NSD1* "negative". These refer to constitutional mutations only. For the purposes of our discussion, *EZH2* positive individuals are considered Weaver syndrome (WS) cases and *NSD1* positive individuals are Sotos syndrome (SS) cases. Frequencies of each trait observed in other *EZH2* positive WS cases reported in the literature were also given when available.

| | **Prevalence in individuals with *NSD1* mutations *** | **Prevalence in individuals with *EZH2* mutations *** | **Prevalence in individuals with *EZH2* mutations as reported in the literature *** [55,58,60] | **Prevalence in *EZH2* and *NSD1* "negative" individuals in our cohort *** |
|---|---|---|---|---|
| Total number | 7 | 11 | 50 | 48 |
| Sex distribution (males; females) | 4 (57%); 3 (43%) | 7 (64%); 4 (36%) | 21 (42%); 29 (58%) | 30 (62.5%); 18 (37.5%) |
| **Growth features** | | | | |
| Excessive growth of PREnatal onset | 2/2 | 11/11 | 22/30 (73%) | 24/37 (65%) |
| Excessive growth of POSTnatal onset | 0/2 | 0/11 | 7/30 (23%) | 9/37 (24%) |
| Tall stature | 5/6 | 10/11 | 43/47 (92%) | 33/45 (73%) |
| Obesity | 1/5 | 6/11 | 1/1 | 13/36 (36%) |
| Accelerated osseous maturation | 4/4 | 8/8 | 26/26 (100%) | 22/31 (71%) |
| **Neurological features** | | | | |
| Hypertonia | 0/4 | 5/11 | 12/40 (30%) | 4/35 (11%) |
| Hypotonia | 2/4 | 7/10 | 18/41 (44%) | 22/37 (60%) |
| Hoarse low-pitched cry | 1/1 | 6/9 | 10/27 (37%) | 7/13 (54%) |
| Intellectual disability | 1/3 | 8/8 | 37/45 (82%) | 27/29 (93%) |
| Excessive appetite | 0/3 | 4/8 | - | 12/31 (39%) |
| Developmental Delay | 4/5 | 5/9 | 0/1 | 39/43 (91%) |
| Speech Delay | 4/4 | 10/10 | 1/1 | 35/39 (90%) |
| Autism spectrum disorder | 1/4 | 0/7 | - | 14/39 (36%) |
| Other behavioural and/or mental disorders | 0/4 | 4/7 | 5/48 (10%) | 30/39 (77%) |
| Ventriculomegaly | 0/1 | 1/8 | 5/48 (10%) | 5/28 (18%) |
| Delayed myelination | 0/1 | 1/5 | - | 2/26 (8%) |
| Cerebellar hypoplasia | 0/1 | 1/8 | - | 1/27 (4%) |
| Seizures | 0/3 | 4/9 | 3/48 (6%) | 11/37 (30%) |
| Polymicrogyria | 0/1 | 1/7 | 2/50 (4%) | 0/27 (0%) |

| | **Prevalence in individuals with *NSD1* mutations *** | **Prevalence in individuals with *EZH2* mutations *** | **Prevalence in individuals with *EZH2* mutations as reported in the literature *** [55,58,60] | **Prevalence in *EZH2* and *NSD1* "negative" individuals in our cohort *** |
|---|---|---|---|---|
| Pachygyria | 0/1 | 1/8 | 1/50 (2%) | 1/27 (4%) |
| Periventricular leukomalacia | 0/1 | 0/8 | 2/50 (4%) | 3/35 (9%) |
| Other brain abnormalities | 0/1 | 5/8 | 1/50 (2%) | 18/35 (51%) |
| Poor fine motor coordination | 2/3 | 9/10 | 28/35 (80%) | 29/35 (83%) |
| Poor balance/gravitational insecurity | 0/1 | 7/8 | - | 21/30 (70%) |
| **Craniofacial** | | | | |
| Macrocephaly | 7/7 | 9/11 | 23/44 (52%) | 33/45 (73%) |
| Rounded head/face (in early years) | 3/7 | 10/11 | 2/2 | 22/40 (55%) |
| Large bifrontal diameter | 2/6 | 10/11 | 2/2 | 18/36 (50%) |
| Prominent forehead | 5/7 | 9/11 | 0/1 | 32/41 (78%) |
| Flat occiput | 1/3 | 8/10 | 1/1 | 9/23 (39%) |
| Large ears ** | 5/7 | 9/11 | - | 12/44 (27%) |
| Otitis media (recurrent) | - | - | - | 11/32 (34%) |
| Ocular hypertelorism | 5/6 | 8/11 | 2/2 | 25/38 (66%) |
| Strabismus | 0/4 | 0/1 | 4/48 (8%) | 6/24 (25%) |
| Myopia | 0/4 | 1/10 | - | 7/39 (18%) |
| Other eye abnormalities | 0/4 | - | - | 15/39 (39%) |
| Down slanted palpebral fissures | 7/7 | 8/10 | 1/1 | 16/42 (38%) |
| Full/thick eyebrows | 0/5 | 3/10 | - | 5/39 (13%) |
| Sparse eyebrows | - | 4/10 | - | 14/39 (36%) |
| Long philtrum | 2/5 | 8/11 | 0/1 | 15/41 (37%) |
| Broad nose | 0/6 | 0/10 | - | 6/40 (15%) |
| Wide nasal root | 0/5 | 7/10 | - | 21/38 (55%) |
| High arched palate | 2/5 | 2/3 | - | 13/28 (46%) |
| Cleft palate | - | - | 3/48 (6%) | - |
| Prominent chin/jaw | 5/5 | 6/11 | - | 16/40 (40%) |
| Micro/retrognathia | 3/5 | 10/11 | 2/2 | 20/39 (51%) |
| Rosy cheeks or malar flushing | 3/6 | 3/10 | - | 12/36 (33%) |
| **Cardiovascular** | | | | |
| Patent ductus arteriosus | 0/1 | 2/8 | 0/48 (0%) | 1/9 |
| Ventricular septal defect | 0/2 | 1/6 | 2/48 (4.2%) | 3/9 |
| Respiratory problems | 2/4 | 3/4 | - | 13/30 (43.3%) |
| **Limbs** | | | | |
| Limited elbow & knee extension in early life | 0/5 | 5/10 | - | 7/28 (25%) |
| Limited elbow & knee extension after puberty | 0/1 | 2/4 | - | 4/6 |
| Widened distal femurs and ulnas | - | 2/4 | - | 1/5 |
| **Hands** | | | | |
| Large hands ** | 4/4 | 4/6 | - | 11/26 (42%) |
| Long slender fingers | 0/3 | 0/6 | - | 7/25 (28%) |
| Prominent digit pads | 1/2 | 8/10 | - | 9/33 (27%) |
| Single transverse palmar crease | 0/3 | 4/10 | - | 5/27 (19%) |
| Camptodactyly | 0/5 | 2/9 | 17/38 (45%) | 6/33 (18%) |
| Clinodactyly | 1/6 | 1/8 | 9/48 (19%) *** | 4/31 (13%) |
| Broad thumbs | 0/3 | 3/9 | - | 9/31 (29%) |
| Thin, deep-set nails | 0/1 | 6/10 | - | 6/34 (18%) |

| | Prevalence in individuals with *NSD1* mutations * | Prevalence in individuals with *EZH2* mutations * | Prevalence in individuals with *EZH2* mutations as reported in the literature * [55,58,60] | Prevalence in *EZH2* and *NSD1* "negative" individuals in our cohort * |
|---|---|---|---|---|
| **Feet** | | | | |
| Large feet ** | 4/5 | 3/4 | - | 7/23 (30%) |
| Clinodactyly (usually 4/5) | 0/6 | 3/10 | - | 10/34 (29%) |
| Syndactyly (2/3) | 2/6 | 0/10 | 1/48 (2%) | 1/34 (3%) |
| Talipes equinovarus | 0/3 | 3/9 | 6/48 (13%) | 3/29 (10%) |
| Short fourth metatarsals | 0/3 | 1/7 | - | 3/29 (10%) |
| Hind foot valgus | 1/4 | 1/8 | - | 9/30 (30%) |
| **Skin** | | | | |
| Excessive loose or "doughy" skin | 1/6 | 7/11 | 17/35 (49%) | 5/38 (13%) |
| Hypoplastic/supernumerary nipples | 2/2 | 5/11 | - | 11/32 (34%) |
| Thin hair | 2/3 | 3/10 | - | 6/34 (18%) |
| Increased pigmented nevi or other marks | 1/4 | 4/10 | 6/48 (13%) | 20/36 (56%) |
| **Connective tissue** | | | | |
| Umbilical hernia | 1/1 | 9/10 | 18/41 (44%) | 6/25 (24%) |
| Inguinal hernia | 0/1 | 1/9 | 0/1 | 2/25 (8%) |
| Diastasis recti | 0/3 | 4/10 | - | 2/29 (7%) |
| Scoliosis | 2/5 | 5/9 | 7/48 (15%) | 9/31 (29%) |
| Kyphosis | 0/5 | 2/10 | 1/49 (2%) | 2/31 (7%) |
| Pectus excavatum | 2/5 | - | 3/48 (6%) | - |
| Pectus carinatum | 0/5 | - | 1/48 (2%) | - |
| **Endocrine** | | | | |
| Hypothyroidism | 0/3 | 1/8 | - | 3/24 (13%) |
| Growth hormone deficiency | 0/2 | 1/7 | - | 1/17 (6%) |
| Hypoglycemia | 1/2 | 3/8 | - | 4/23 (17%) |
| **Neoplasia** | total 1/7 | total 3/11 | total 3/50 [x] (6%) | total 0/48 |
| Neuroblastoma | 1/3 | 2/10 | 1/50 (2%) | 0/31 (0%) |
| Leukemia | 0/3 | 1/11 | 2/50 (4%) | 0/32 (0%) |
| Lymphoma | 0/3 | 0/11 | 1/50 (2%) | 0/32 (0%) |
| **Other** | | | | |
| Cryptorchidism | 1/4 | 1/4 | 1/48 (2%) | 4/18 (22%) |
| Early puberty | - | 1/4 | - | 5/16 (31%) |

**Table 2-7: Comparing the prevalence of phenotypic features between mutation positive and mutation negative individuals.**

* Not every feature was described for every patient thus only informative numbers are given for each feature; percentages are only given for features informative for twelve or more individuals in the larger cohorts (N=48-50). Further, this table refers to presence or absence of each feature in each patient and does not reflect prominence of the feature (present included mild to severe/very prominent). Characteristics that were described as present then resolved at a later date were also counted as present. Unspecific or unclear descriptions were considered non-informative. Individuals carrying likely pathogenic variants were considering mutation positive for this table.

** Not specified if proportional or not to generalized overgrowth. *** Not specified if clinodactyly of digits or toes.

[x] There were four different tumors reported, but two were in the same individual. - = non informative.

#### 2.4.1.1 Weaver vs. Sotos

Significant phenotypic differences between *EZH2* positive and *NSD1* positive individuals from our cohort cannot be determined because of the small sample sizes (N=11 and N=7). Moreover, a similar comparison between clinical features of WS and SS has already been carried out in a larger study by Tatton-Brown and Rahman (although exact sample sizes were not provided).[61] In general, findings within our own cohort were consistent with previous observations.

It is worth noting that the prevalence of cancers among mutation-positive individuals was estimated by Tatton-Brown and Rahman to be around 3% for *NSD1* mutation positive individuals and 5% for *EZH2* mutation positive individuals.[61] Within our cohort, we observed a cancer prevalence of 14.3% (or 1/7) for SS and 27.3% (or 3/11) for WS. Both values are clearly over-estimates due to ascertainment bias within this highly selective cohort; nonetheless, these observations are in alignment with the number of somatic mutations described in the COSMIC database (470 coding mutations in *NSD1* versus 821 coding mutations in *EZH2*), supporting that the risk for cancer development in WS is likely higher than in SS. Importantly, all three malignancies observed in our *EZH2* positive cohort developed by the age of 11 months, and the malignancy observed in the *NSD1* positive patient was detected before the age of 3 years. This suggests that cancer prevalence in WS and SS may be higher in early years, similar to what is observed in other overgrowth disorders such as Beckwith-Wiedemann syndrome (see Chapter 1, section 1.1.3). Furthermore, of the twelve malignancies reported to date in WS patients (including our own, six with molecularly confirmed *EZH2* mutations), half of them were diagnosed by age 13 months, two others around age 4 years, and the remaining four were detected during adolescence (13-19 years).[58–60] Based on these observations, we recommend a tumour surveillance program for WS patients with clinically-validated *EZH2* mutations. This screening should include physical examinations, blood tests and imaging to look for signs consistent with blood-derived cancers and neuroblastomas, as these appear to be the most common tumour types in WS. These tests should be more frequent in early years (every 2-3 months in the first 4 years of life), and subsequently from puberty until early adulthood (between 12 and 20 years).

#### 2.4.1.2 Prevalence of Weaver-like features in our *EZH2* positive cohort in relation to other reported cases

We identified eleven *EZH2* positive individuals within our cohort; for one of these individuals (proband 11), the inheritance mode of the variant has not yet been determined. There are now 50 other cases reported in the literature (which form the "literature" cohort or LC), the majority of which (48/50) are described by Tatton-Brown *et al*. in a single article.[58] It should be noted that the phenotypic information published for these cases was limited, and that the inheritance of the variants was not established for 13/48 cases, including novel variants for which pathogenicity is therefore not confirmed. Nonetheless, these cases can help us understand the phenotypic spectrum of *EZH2* positive individuals.

In general, the overall phenotypic presentation of WS cases was very similar between our study (N=11) and the LC (N=50), as expected. Some features showed a higher calculated prevalence in our cohort; these included: behavioural disorders, seizures, brain abnormalities, cardiac defects, spinal deformities, and cancer. These perceived enrichments are due to ascertainment bias, with clear differences between the two cohorts. Indeed, in our study we have collected mostly sporadic cases with severe phenotypes; in contrast, many of the cases described in Tatton-Brown *et al*.[58] are familial, and thus possibly recruited based on having several affected members within the same family and not relying as heavily on phenotypic severity. These numbers also illustrate the need to assemble larger cohorts (ideally followed longitudinally) in order to accurately determine the prevalence each phenotypic trait in WS.

#### 2.4.1.3 Expanding the Weaver phenotype to include neuronal migration disorders

Brain imaging in proband 5 (Figure 2-2g) was consistent with that reported on two prior occasions in different children with WS.[40,55] In each previous case, and in our case, there was asymmetric perisylvian polymicrogyria that appeared more severe on the right side, as well as mildly enlarged lateral ventricles. The report by Freeman *et al.*[40] described pachygyria, but based on review of the published images (W. Dobyns), we believe the findings are more consistent with perisylvian polymicrogyria. The image shown in the report by Al-Salem *et al.*[55] demonstrates enlarged extra-axial fluid spaces over the brain, similar in appearance to the polymicrogyria observed among megalencephaly syndromes associated with PI3K-AKT pathway mutations.[288] However, neither hydrocephalus nor Chiari malformations were seen in

these WS patients. Tatton-Brown *et al.*[58] also reported a case with pachy- and polymicrogyria, but no images were published.

Our proband 5 has the recurrent p.(Arg684Cys) *de novo* mutation, and the patient with polymicrogyria and WS described by Al-Salem *et al.*[55] was shown to have a *de novo* p.(Glu745Lys) mutation. The patient from Tatton-Brown *et al.*[58] had an *EZH2* variant predicted to truncate the protein at position 732, classified as likely pathogenic because parental samples were unavailable. The association of polymicrogyria with WS in four independent cases, two of which have molecular confirmation of *de novo* mutations in different exons of *EZH2* and another of which has a truncating variant in the last exon, strongly support a causal association between *EZH2* mutations and neuronal migration defects in some patients with WS. EZH2 has been shown to control the decision between self-renewal and differentiation in the cerebral cortex, and inhibition of PRC2 complex activity had been shown to shift the balance toward differentiation.[289] Furthermore, EZH2 has also been shown to orchestrate neuronal migration in the cortico-ponto-cerebellar pathway in mice.[290] The possibility that diminished PRC2 complex activity could lead to premature neuronal differentiation, possibly at ectopic sites along the normal migration pathway, offers a plausible explanation for the cortical patterning defects seen in patients with WS and pachy- or polymicrogyria.

Thus, in addition to their known risk for neoplastic disease, patients with WS should be considered to be at risk for neuronal migration disorders, and physicians should have a low threshold for ordering cranial imaging studies. Similarly, children with overgrowth and cerebral migration disorders could be tested for rare variants in *EZH2*, and physicians performing prenatal diagnosis in the context of a fetus with polymicrogyria should consider the possibility of WS. Although the finding of a neuronal migration disorder in a child with WS may not necessarily change their clinical management, it may assist in counselling the parents with regards their child's day-to-day care, and in determining the additional assistance that may be required. Furthermore, the knowledge that *EZH2* mutations can cause neuronal migration disorders may allow for faster diagnosis of new cases of polymicrogyria, which in turn would inform them on their prognosis in the context of WS.

Given the large number of individuals now being studied with high-throughput next-generation sequencing, rare variants in *EZH2* that are discovered through targeted sequencing panels, exome sequencing or whole genome sequencing should be considered carefully in the

context of clinical findings such as overgrowth phenotypes, cerebral malformations and neoplastic disease. For highly heterogeneous disorders like overgrowth syndromes, techniques like exome sequencing are becoming cost-effective for diagnosis at an early stage of the workup. An estimate of diagnostic costs incurred during the workup of proband 5 is presented in Appendix F, for theoretical comparison to early exome sequencing (though in his particular case, *EZH2* was selected as a candidate for screening on the basis of his facial dysmorphism as a toddler).

#### 2.4.1.4 The *EZH2*/*NSD1* negative cohort is likely to represent a heterogeneous group of disorders

Within the *EZH2*/*NSD1* mutation negative cohort (NC) described in Table 2-7, intellectual disability, developmental and speech delay, and poor fine motor coordination are the most common features. This is not surprising because these traits are shared by a large number of developmental disorders. Furthermore, no feature appears to be particularly common or uncommon, most ranging between 25 and 75% in frequency, suggesting that this is a phenotypically heterogeneous cohort. This variability is likely increased by the fact that the cohort contains a mixture of children and adults at various ages, and that many individuals were referred by the families themselves rather than by physicians.

Phenotypic heterogeneity in the NC strongly suggests genetic heterogeneity in the underlying causative genes. Some individuals may actually have alterations in *EZH2* or *NSD1* that were missed by Sanger sequencing (as discussed later, section 2.4.3), while others will have a myriad of genetic and genomic alterations involving other genes. Attaining a concrete molecular diagnosis for these patients will be a challenge, as has been previously recognized for rare disorders in general [291,292] and in particular for individuals with intellectual disability.[293,294] Grouping unrelated cases with the most phenotypic overlap, as was successfully done by Gibson *et al.*[56] and many others, may increase the chances of identifying new causal genes; however, successful gene discovery will not be guaranteed due to the phenotypic variability observed even within most syndromes. Photographs to determine specific facial gestalt (ideally from early years and at approximately the same age across patients) will be most helpful.

### 2.4.2 Clarifying the mutational spectrum of *EZH2* in Weaver syndrome

As mentioned previously, pathogenic mutations identified in *EZH2* are scattered throughout the gene (see Figure 2-1 for our cohort-specific variants, or Figure 1-1 for all variants). However, there does appear to be some clustering within exon 5 and within the SET domain, as discussed below. All mutations identified within our cohort are missense except for one indel (small deletion), and some are recurring.

Probands 1, 2, 4, 6, 7, 8, and 11 (total: 7/11) all have mutations within exon 5. p.(Pro132Ser) and p.(Tyr133Cys) were each identified in two unrelated individuals. Of interest, a p.(Pro132=) synonymous variant has been reported at low frequency in normal populations (rs61732845). One patient with the p.(Pro132Ser) mutation and one with the p.(Tyr133Cys) mutation developed neuroblastomas extremely early, and a p.(Pro132Leu) mutation was recently reported in a WS case that developed acute myeloid leukemia.[60] Variants affecting these amino acids (including p.Pro132Ser) have also been described in somatic cancers (see Appendix G). These findings are interesting because the majority of exon 5 does not correspond to a recognizable functional domain, yet the increased number of disease-associated variants reported within this exon may point toward an as-yet unappreciated functional role specific to exon 5.

The mutations identified in probands 3, 5 and 9 (total 3/11) are located within the SET domain, which is known to be the catalytic domain of EZH2, and so it is not surprising that mutations within this domain would lead to disease. In fact, cancer "hotspot" mutations affect the tyrosine residue at position 646 which is located within the SET domain, and represents the most frequently mutated amino acid in *EZH2*. The p.(Arg684Cys) mutation identified in proband 5 is recurrent, as it has been described in an additional five WS patients.[58] Variants affecting amino acids within the SET domain, including p.(Arg684Cys) and p.(His694Tyr), have also been described in somatic cancers (see Appendix G).

The p.(Glu745Lys) mutation identified in proband 10 is located in exon 20, between the SET domain and the C-terminus. Its proximity to the end of the protein (which is 751 amino acids long) may call its pathogenicity into question. However, this mutation was found to be *de novo* in this patient, as well as in another patient described by Tatton-Brown *et al*.[58], and both developed haematological malignancies. This variant has also been described in somatic cancers (see Appendix G).

Together, our results show that alterations across *EZH2*, and particularly within exon 5 and the SET domain, can lead to a WS phenotype. It is unclear whether mutations in other domains have not been reported because they are extremely rare, because they are associated with a milder phenotype that has not been ascertained, or because they are associated with a much more severe phenotype that might not be tolerated (and cause embryonic lethality). It is also unclear if certain mutations are associated with a more severe phenotype (for example due to a more significant impairment in protein function); the work described in Chapter 3 partially addresses this question. However, individuals with the same mutation often do not present with the same phenotypic severity, suggesting that knowing which *EZH2* mutation is causing disease is not sufficient to infer the clinical impact of the mutation; this is particularly important in the context of prenatal testing, where a decision of pregnancy termination could be made based on variant interpretation.

### 2.4.3 Diagnostic rates by sequencing known overgrowth genes

The complete diagnostic progression is summarized in Chapter 5 (see Figure 5-1).

*EZH2* sequencing was carried out for 54 individuals in the overgrowth cohort. Validation of external results was carried out for a further eight individuals. Overall, 10/66 probands (equivalent to 15% of the total cohort) had confirmed pathogenic mutations and were classified as positive for *EZH2*, with a recommended molecular diagnosis of Weaver syndrome (WS). Another proband has a likely pathogenic variant and is awaiting testing of parental samples for definitive classification, which could increase the diagnostic rate to 17% (or 11/66). However, if we exclude the three original cases from Gibson *et al.*,[56] which had been described by specialists as having "classical" WS and thus had a higher prior probability of testing positive, our detection rate for *EZH2* mutations is actually 13% (or 8/63). This rate is comparable to that described by Tatton-Brown *et al.*,[58] who identified *EZH2* mutations in 48/435 individuals (11%) in their wider overgrowth cohort. All families and/or referring physicians were encouraged to seek validation of results in a clinically certified laboratory so that results could potentially be used for clinical management.

For individuals who tested negative for rare coding mutations in *EZH2* and had no alternative molecular diagnosis, *NSD1* sequencing was carried out only when the phenotype was consistent with SS. A total of 20 individuals were sequenced, and validation of external results

was carried out for a further five individuals. Overall, 5/66 probands (equivalent to 8% of the total cohort, or 9% of the remaining undiagnosed patients) had pathogenic mutations and were classified as positive for *NSD1*, with a recommended molecular diagnosis of Sotos syndrome. A further two probands carry likely pathogenic variants, which could increase the diagnostic rate to 11%. Again, families were encouraged to seek clinical validation of these results.

Together, 15/66 probands were successfully diagnosed using Sanger sequencing of *EZH2* and *NSD1*. This corresponds to a 23% molecular diagnostic rate (or 27% if we include the likely pathogenic variants observed in an additional three probands). This rate is considerably high given that we only carried out targeted gene testing for two genes, and especially if we consider the diagnostic rate within the Weaver-like cohort only (18/47, or 38%). Furthermore, if we add the two probands who were found to have an alternative diagnosis through independent investigations, the diagnostic rate within our full cohort increases to 30%. This rate supports the utility of collecting extremely detailed phenotypic information to aid in molecular diagnosis, although it also illustrates that the detection rate for *EZH2* and *NSD1* mutations within a cohort ascertained for generalized overgrowth and dysmorphic features is overall low, leaving many patients undiagnosed after candidate gene testing (unlike other conditions such as cystic fibrosis).

It is important to note that Sanger sequencing only detects single nucleotide variants (SNVs) and small insertions or deletions (indels); other types of genetic alterations such as structural variants, copy number variants (CNVs) and trinucleotide expansion repeats are not detected. Although most probands have had negative microarray results that exclude large deletions or duplications at these loci, smaller CNVs (encompassing a single exon for example) may be present. Microdeletions have been previously reported in *NSD1* (see Chapter 1, section 1.1.2) and could exist in *EZH2*, but these require other methods such as MLPA (Multiplex Ligation-dependent Probe Amplification) for detection. Furthermore, alterations at the RNA level have not been interrogated. Recent studies have shown that some seemingly benign variants identified in known disease genes (within both exons and introns) may actually affect RNA splicing through unconventional mechanisms.[295,296] These "likely benign" variants include coding synonymous variants,[297–299] meaning that a subset of variants previously interpreted as unlikely to cause disease may actually be pathogenic. This additional variation would also be missed upon exome sequencing, particularly when considering non-coding regions. As such, individuals

referred here as being negative for *EZH2* or *NSD1*, are only negative for clearly pathogenic SNVs and small indels within or near the coding regions of these genes, and could in fact have other alterations within these genes. These methodology limitations are important on a diagnostic level, as physicians may be excluding compelling candidate genes based on incomplete results, and also on a functional level, because the identification of other types of alterations within a gene could provide new hypotheses with regard to the molecular mechanism of disease.

# Chapter 3: *In vitro* studies suggest that Weaver syndrome is caused by an impairment in EZH2-mediated histone methyltransferase activity

## 3.1 Introduction

### 3.1.1 EZH2 function in cancer and Weaver syndrome

The histone methyltransferase EZH2 is a key chromatin regulator in mammals. This protein forms the catalytic subunit of the Polycomb Repressive Complex 2 (PRC2),[178] and is thought to suppress gene transcription by adding up to three methyl groups onto lysine residue 27 of histone H3 (H3K27), which influence chromatin state (see Chapter 1).

Somatic mutations of *EZH2* in circulating white blood cells have been shown to be extremely common in haematological malignancies.[229] Sequencing of human diffuse large B-cell and non-Hodgkin lymphomas revealed recurrent somatic mutations at positions Tyr646, Ala682 and Ala692 (or Tyr641, Ala677 and Ala687 in the shorter EZH2 isoform as referenced in the original publications).[254,258,259] The most frequently mutated residue, tyrosine at position 646, has been reported as mutated to phenylalanine (p.Tyr646Phe), asparagine (p.Tyr646Asn), histidine (p.Tyr646His) and serine (p.Tyr646Ser).[254] All of these heterozygous single amino acid substitutions have been shown to alter substrate specificity and favour trimethylation of H3K27 (H3K27me2 → H3K27me3) over monomethylation (H3K27me0 → H3K27me1) and dimethylation (H3K27me1→ H3K27me2). When combined with the activity of the wild-type (WT) EZH2 copy, which has high affinity for unmethylated H3K27me0 and medium affinity for monomethylated H3K27me1 but low affinity for dimethylated H3K27me2, this results in overall gain-of-function of EZH2.[190,257] Unlike for Tyr646, mutations at the other two sites have only been reported as specific amino acid substitutions, (p.Ala682Gly)[258] and (p.Ala692Val).[259] These two alterations also show gain-of-function activity but appear to favour different substrates. p.Ala682Gly promotes methyltransferase activity of nearly equal efficiency for all three substrates (H3K27me0/me1/me2), thereby resulting in hypertrimethylation of H3K27.[258] By contrast, p.Ala692Val reduces monomethylation and enhances dimethylation of H3K27 while leaving trimethylation virtually unchanged *in vitro*,[259] though global levels of H3K27me3 were found to be increased in a tumour-derived p.(Ala692Val) mutant cell line, or when this mutant was transiently expressed in a cell line with an EZH2 WT background.[300] Given that additional

data support the increased activity of EZH2 in other cancers, possibly independent of its assembly into PRC2 (see Chapter 1, section 1.2.1.2.1),[251] there is intense interest in developing EZH2 inhibitors as potential chemotherapeutic agents.[260–263,301,302] Targeting of other proteins in the complex, such as EED and SUZ12, may also become a useful therapeutic strategy,[301] as may disruption of proper PRC2 assembly.[303] Notably, though, inactivating mutations have been found at multiple sites throughout the *EZH2* gene in myeloid disorders and acute lymphoblastic leukemias,[231–233,264,265,304,305] suggesting that some neoplasms may not respond well to EZH2 inhibition.

We and others have shown that *de novo* germline mutations in *EZH2* cause Weaver syndrome (WS), a rare but well-described developmental disorder of prenatal onset that features intellectual disability, tall stature, macrocephaly, accelerated bone growth and maturation and a susceptibility to cancers including haematological malignancies.[56–58,61] Aspects of this phenotype can be explained by the role of Ezh2 in craniofacial skeleton formation.[306] Though cerebral malformations were not part of the original description of WS,[10] recent clinical reports have documented the presence of neuronal migration disorders in association with physical features of WS,[40,55,58] similarly to what we observed in proband 5 (see Chapter 2, section 2.4.1.3).

Given that both gain-of-function and loss-of-function mutations in *EZH2* have been associated with human neoplastic disease when acquired during life in somatic cells,[229,232,257,258,264] we hypothesized that germline *de novo* mutations causative of WS would alter EZH2 activity within PRC2. We further hypothesized that more severe clinical features of WS (such as cerebral migration defects, or the development of leukemia) might be specifically associated with more significant changes in PRC2-mediated methyltransferase activity conferred by individual mutations in *EZH2*.

### 3.1.2 Mechanism of disease for NSD1 in Sotos syndrome

Sotos syndrome (SS) is caused by mutations [46,130–133] or microdeletions [129] within *NSD1*, which encodes an H3K36 histone methyltransferase. Thus SS and WS may be both phenotypically and functionally similar. Due to the nature of the alterations in *NSD1*, which included numerous truncating mutations, a haploinsufficiency mechanism was inferred in 2002,[128] although no functional studies were carried until much later. In 2011, Qiao *et al*.[277] showed that five different *NSD1* missense mutations that had been identified in SS patients lead

to reduced histone methyltransferase activity *in vitro*, ranging from a partial reduction to undetectable levels of histone methyltransferase activity. These findings were later confirmed and expanded by Kudithipudi *et al*.,[307] who formally suggested chromatin state modification as the mechanism for the pathogenesis of SS. However, we must remember that methylation of H3K36 (by NSD1) leads to gene activation, whereas methylation of H3K27 (by EZH2) leads to gene repression, and that the genomic targets of these methylation reactions are different. Thus, it is important to test for the function of EZH2 in WS separately.

## 3.2 Methods

To test our hypothesis, we designed recombinant human EZH2 proteins, had them preassembled into complexes together with other PRC2 proteins (with BPS Bioscience), and tested their activity *in vitro* using a well-accepted *in vitro* assay.[231,257,264] Mutant EZH2-PRC2 complexes were selected for study based on the rare *de novo* variants observed in our original three patients [56] and among other patients with WS identified since. We also tested a common variant detected in multiple individuals within our cohort. Sequencing results for these patients are described in Chapter 2 section 2.3.2.2 and summarized in Table 3-1 below.

| **Patient** | ***EZH2* variant** | **Inheritance** | **Interpretation** |
|---|---|---|---|
| Proband 1 | p.(Tyr153del) | *de novo* | pathogenic |
| Proband 2 | p.(His694Tyr) | *de novo* | pathogenic |
| Proband 3 | p.(Pro132Ser) | *de novo* | pathogenic |
| Proband 4 | p.(Pro132Ser) | *de novo* | pathogenic |
| Proband 5 | p.(Arg684Cys) | *de novo* | pathogenic |
| Proband 6 | p.(Tyr133Cys) | *de novo* | pathogenic |
| Proband 7 | p.(Tyr133Cys) | *de novo* | pathogenic |
| Case 15 | p.(Asp185His) | paternally inherited | likely benign |
| Case 40 | | paternally inherited | |
| Case 53 | | maternally inherited | |
| Case 73 | | maternally inherited | |
| Case 95 | | maternally inherited | |
| Cases 118/119 | | paternally inherited | |
| Case 130 | | paternally inherited | |

**Table 3-1: Summary of *EZH2* variants identified in our cohort and tested via *in vitro* functional assays.**

Further, two other variants that had been observed in patients with both WS and neoplastic disease [57] were selected. One of these variants, p.(Glu745Lys), was later observed in proband 10 in our cohort, who also presented with both WS and early onset cancer (see Chapter 2, section 2.3.2.2.1). In addition, wild-type (WT) EZH2 was used as a positive control, and the methyltransferase-inactive mutant EZH2 p.Phe672Ile (equivalent to the inactive fly mutant allele $E(z)^{son1}$ described by Joshi *et al.*[308]) was used as a negative control (see Table 3-2 below).

| ***EZH2* variant** | **Reason for testing** | **Predicted function** | **Reference** |
|---|---|---|---|
| Wild-type | Reaction control | Normal function: high mono- and di-methylation, low trimethylation | [257] |
| p.(Phe672Ile) | Negative control – inactive *Drosophila* mutant | No function – dead enzyme | [308] |
| p.(Tyr646Ser) | Positive control – hyperactive cancer mutant | Hypertrimethylation (although known to have reduced function against mixed histones) | [190,257] |
| p.(Ala682Thr) | Neoplasm testing - identified in patient with WS features who developed a neuroblastoma and acute lymphoblastic leukemia | Inactivating | [57] |
| p.(Glu745Lys) | Neoplasm testing - identified in patient with WS features and lymphoma, also later identified in proband 10 from our cohort who has WS features and acute lymphoblastic leukemia | Inactivating | [57] |

**Table 3-2: Additional *EZH2* variants tested via *in vitro* functional assays.**

### 3.2.1 Assay materials

Core histones were purchased from Millipore (13-107) and used as methyl acceptors in most of our *in vitro* reactions. This substrate contains a mix of all core histones and H3 peptides at different methylation states, thereby recapitulating the heterogeneity of endogenous nucleosomes. In an alternative assay, biotinylated peptides (mimicking the H3 tail, H3(21-44)) that had been unmethylated (H3K27me0), monomethylated (H3K27me1) or dimethylated (H3K27me2) were used as substrate (Figure 3-4). We purchased PRC2 complexes containing wild type EZH2 (#51004) or mutant EZH2 from BPS Bioscience.[254] Methyltransferase assays were done using a commonly-used kit[231,257,264] (17-330, Millipore) as per manufacturer's instructions.

### 3.2.2 Optimization of assay conditions

This assay was optimized based on the protocol described in Yap, *et al*.[257] A schematic representation of the enzymatic reaction being assayed is presented in Figure 3-1 below.

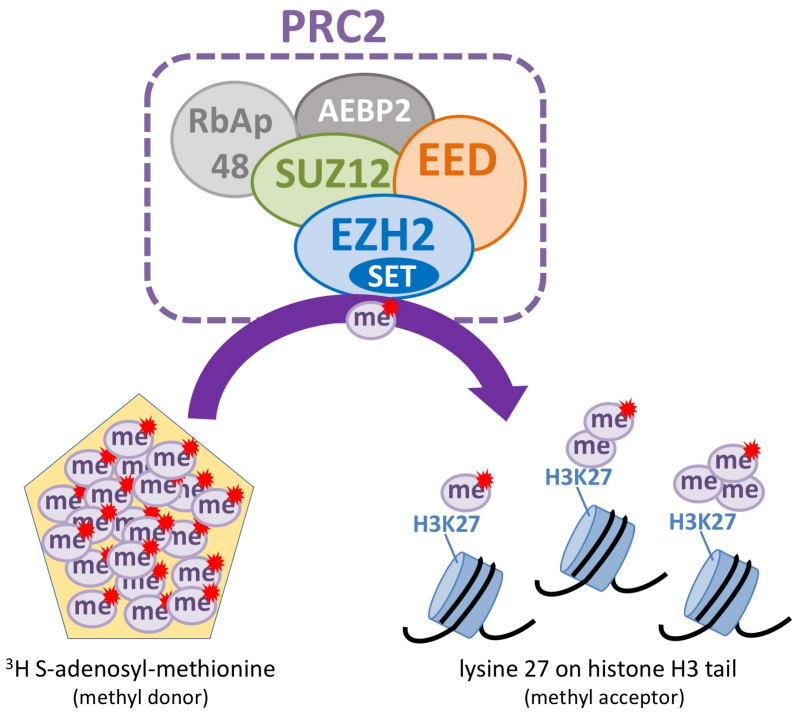



**Figure 3-1: Schematic representation of the histone methyltransferase reaction measured by our *in vitro* assay.**

Human WT or mutant EZH2 is pre-assembled into the PRC2 complex. EZH2 is the catalytic member of PRC2, capable of transferring methyl groups (me) from a methyl donor (here the radioactively labeled S-adenosyl-methionine or $^3$H-SAM) to a methyl acceptor (here the purified core histones at mixed methylated states). EZH2-PRC2 can add up to three methyl groups to a single H3K27 residue. Methyl groups that are not bound to H3K27 are washed away. Histone methyltransferase activity can be measured by detecting the level of radioactivity from the remaining $^3$H labeled methyl groups.

First, the incorporation of tritiated methyl groups from $^3$H-labeled S-adenosyl-methionine (SAM) onto core histones was measured in the presence of either G9a, wild-type EZH2-PRC2 complex or each of three mutant forms of the EZH2-PRC2 complex (Figure 3-2). G9a was used as a positive control enzyme for histone methyltransferase activity, given that this enzyme is

active towards purified nucleosomes on its own,[309,310] in contrast to EZH2 which requires other members of the PRC2-complex for activity on nucleosomes.[178] As expected, wild-type EZH2-PRC2 complex catalyzed the incorporation of $^{3}$H into core histones, consistent with the model whereby it uses $^{3}$H-SAM as the methyl donor and nucleosomes as the substrate for histone methylation.[178] In contrast, PRC2 complexes containing WS-associated EZH2 p.Tyr153del, p.His694Tyr and p.Pro132Ser mutants showed reduced histone methyltransferase activity (Figure 3-2), suggesting that *EZH2* mutations associated with WS are loss-of-function mutations.

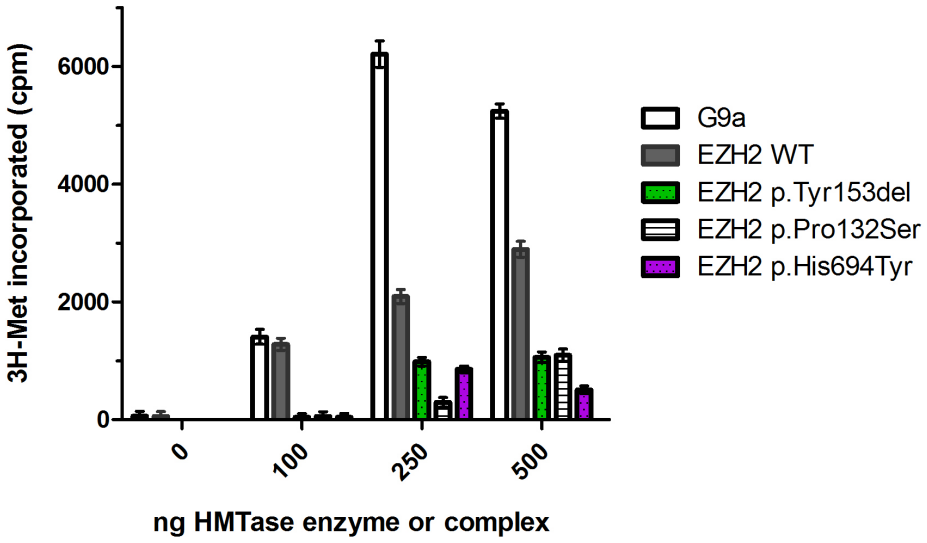



**Figure 3-2: Preliminary data suggesting that EZH2 Weaver syndrome mutants have impaired histone methyltransferase activity *in vitro*.**

Histone methyltransferase reactions were performed using 2µg purified core histones. Individual HMTase complexes were separately incubated with 0.67 µM $^{3}$H-SAM. Reactions were incubated with 0, 50, 100 or 250 ng of either G9a enzyme, which is active on its own (thin solid line), or purified PRC2 complex containing wild-type/WT (up to 500 ng) (thick solid line) or mutant EZH2 p.Pro132Ser (dotted line), p.Tyr153del (dashed line) and p.His694Tyr (dashed-dotted line). Mean incorporation of tritiated methyl groups from $^{3}$H-labeled S-adenosyl-methionine (SAM) onto core histones is shown. Error bars represent calculated standard deviations of replicates from two independent experiments.

Next, a series of experiments was carried out to determine whether the standard concentrations of $^{3}$H-SAM and core histones used, or reaction time, could be limiting the

reaction. Indeed, in order to observe the true enzymatic activity of EZH2, it was important to make sure that all reaction substrates were in excess for all different complexes.[311] Our technical development process is described in detail in Appendix H.

### 3.2.3 Final histone methyltransferase assay

We incubated 250 ng of individual HMTase complexes separately with 0.67 µM $^{3}$H-SAM and 2 µg core histones (or 1 µM peptide), in 50 mM Tris-HCl, pH 9.0 and 0.5 mM DTT for 30 min at 30˚C in a 10 µl volume. In all reactions, five or eight microlitres were spotted on a P81 square paper (Millipore), washed (three times with 10% tricholoracetic acid and once with 95% ethanol) to remove unincorporated $^{3}$H-Met, air-dried overnight, placed in a glass scintillation vial with 3 ml of scintillation fluid (ScintiSafe Econo1 SX20-5 or Scintisafe 30% SX23-5, Fisher Chemical) and counted on a 1900TR Liquid Scintillation Analyzer (Perkin Elmer) or LS6500 Multi-Purpose Scintillation Counter (Beckman Coulter). Normalized counts represent the subtraction of background counts (i.e. control tubes with no enzyme added) from total counts.

## 3.3 Results

### 3.3.1 *EZH2* mutations observed in Weaver syndrome impair histone methyltransferase activity *in vitro*

In order to determine the functional impact of the *EZH2* mutations observed in WS, EZH2 mutant proteins corresponding to the mutations observed in WS patients were expressed *in vitro* and then assembled together with other artificially-expressed members of the PRC2 complex (EED/SUZ12/RbAp48 and AEBP2). Pre-assembly into the complex was necessary because EZH2 requires other members of PRC2 for activity on nucleosomes.[178] The mutations studied included those identified within our cohort: p.(Tyr153del), p.(His694Tyr), p.(Pro132Ser), p.(Arg684Cys), and p.(Tyr133Cys) (see Table 3-1). The p.(Ala682Thr) and p.(Glu745Lys) mutations were also of interest because of their associations with neuroblastoma, acute lymphoblastic leukemia and lymphoma in WS patients (see Table 3-2); we also included the common variant p.(Asp185His) which was detected in multiple individuals within our cohort. Wild-type EZH2 was used as a positive control, and EZH2 p.Phe672Ile as a negative control (see Table 3-2). We then measured incorporation of tritiated methyl groups from $^{3}$H-SAM onto mixed

core histones in the presence of each EZH2-PRC2 complex (Figure 3-3). Wild-type EZH2-PRC2 complex catalyzed the incorporation of $^{3}H$ into core histones as expected, whereas PRC2 complexes containing WS-associated EZH2 mutants showed reduced histone methyltransferase activity *in vitro* (Figure 3-3, Appendix H.1), suggesting that *EZH2* mutations associated with WS are loss-of-function (hypomorphic) mutations.

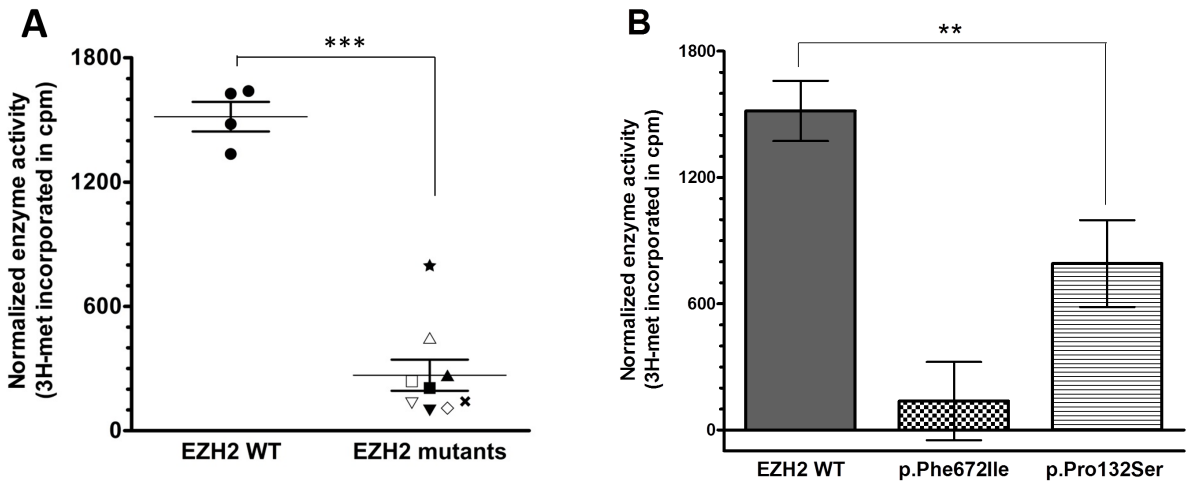



**Figure 3-3: Weaver syndrome mutants are impaired in their histone methyltransferase activity *in vitro*.** Histone methyltransferase reactions were performed using 2 µg purified core histones and 0.67 µM $^{3}H$-SAM. Each reaction was incubated with 250 ng of either wild-type (WT) or a mutant HMTase complex (or no enzyme controls). Histone methyltransferase activity was measured based on the incorporation of $^{3}H$-labeled methyl groups, represented in scintillation counts per minute. Counts were normalized by subtracting background counts (i.e. no enzyme) from the total counts. (A) Incorporation of tritiated methyl groups from $^{3}H$-SAM onto core histones is shown for each complex: EZH2 WT ●, p.Phe672Ile ✖, p.Pro132Ser ★, p.Tyr153del △, p.His694Tyr ▽, p.Glu745Lys ▲, p.Ala682Thr ▼, p.Arg684Cys ■, p.Tyr133Cys □, p.Asp185His ◇. Error bars represent standard deviations within the groups "EZH2 wild-type" and "EZH2 mutants". Unpaired T-Test showed statistically significant difference between the two groups (p-value < 0.0001). (B) Incorporation of tritiated methyl groups from $^{3}H$-SAM onto core histones is shown for the positive control EZH2 WT, the negative control EZH2 p.Phe672Ile and the mutant complex with activity closest to wild type, namely EZH2 p.Pro132Ser. Error bars represent standard deviations of four independent replicates for the controls, and three independent replicates for the mutant EZH2 p.Pro132Ser. One-way ANOVA showed statistically significant difference between all groups (overall p-value < 0.0001; p-values between WT and p.Phe672Ile, between p.Phe672Ile and p.Pro132Ser, and between WT and p.Pro132Ser were all <0.05).

### 3.3.2 The common p.(Asp185His) variant also appears to impair histone methyltransferase activity *in vitro*

The EZH2 p.Asp185His mutant also showed impaired histone methyltransferase activity in this *in vitro* assay (Figure 3-3A, Appendix H.1). Based on the frequency of this variant and the rarity of WS, p.(Asp185His) cannot by itself be causative of WS. However, based on the number of replicates we performed under varied conditions (Figure 3-3A, Appendices H.1 and H.2), we believe this result to be reproducible and to reflect accurately the activity of this enzyme variant under these artificial conditions.

## 3.4 Discussion

### 3.4.1 Weaver syndrome mutations and neoplastic disease

With somatic mutations in *EZH2* having been associated to both gain and loss of histone methyltransferase function, it was important to investigate mutations found in WS patients who had also developed malignancies. To date, none of the three cases from our original report have been diagnosed with neoplastic disease. However, proband 4 had a non-metastatic stage 4S neuroblastoma in his left adrenal gland. Our functional analysis of the EZH2 p.Pro132Ser mutant complex suggested a loss-of-function effect, consistent with a previous report of this variant in the context of myeloid disorders.[305] Proband 6 had a prenatal neuroblastoma in the right adrenal gland that was successfully removed surgically shortly after birth, and the EZH2 p.Tyr133Cys mutant complex also appeared to be loss-of-function in our assay. The other two mutant complexes containing mutations found in patients with malignancies, EZH2 p.Arg682Thr and p.Glu745Lys,[57] also showed loss-of-function *in vitro*, suggesting that the mechanism driving cancer in WS patients resembles that of myeloid disorders and acute lymphoblastic leukemias rather than that of diffuse large B-cell and non-Hodgkin lymphomas. This is also consistent with the development of acute lymphoblastic leukemia in proband 10 of our cohort, who was found to also carry the *de novo* p.(Glu745Lys) mutation after these experiments were carried out. Furthermore, the p.(Arg684Cys) mutation reported in several independent cases [57] and identified in proband 5, which has not yet been associated with malignancy development in WS but

appears to be a true recurrent mutation, had already been described as likely inactivating in myeloid disorders.[264]

Overall, all *de novo* WS-associated *EZH2* mutations showed impaired histone methyltransferase activity *in vitro*, particularly with reduced ability to monomethylate H3K27 (Figure 3-4). Impaired histone methyltransferase activity had previously been observed among *NSD1* mutations causing Sotos syndrome,[277,307] which is another overgrowth syndrome that shares significant phenotypic overlap with WS (as discussed earlier).[61] Based on these results, we suggest that EZH2 inhibitors currently being developed against various cancers[260–263] may not be of benefit in WS. Importantly, we did not assay PRC2-independent functions of EZH2 which have been previously observed in some cancers (see Chapter 1, section 1.2.1.2.1), so additional complexity in the functional effects of disease-associated *EZH2* mutations remains possible. Furthermore, we also did not assay patient-derived samples. Ideally, we would like to measure levels of H3K27me1/2/3 in cells lines derived from patients' lymphocytes and/or fibroblasts, with such cells lines representing a better model of the *in vivo* conditions of these EZH2-PRC2 mutant complexes (with accurate complex assembly and presence of accessory proteins). The exact methods to quantify H3K27me1/2/3 levels would need to be sensitive enough to detect mild changes in methylation.

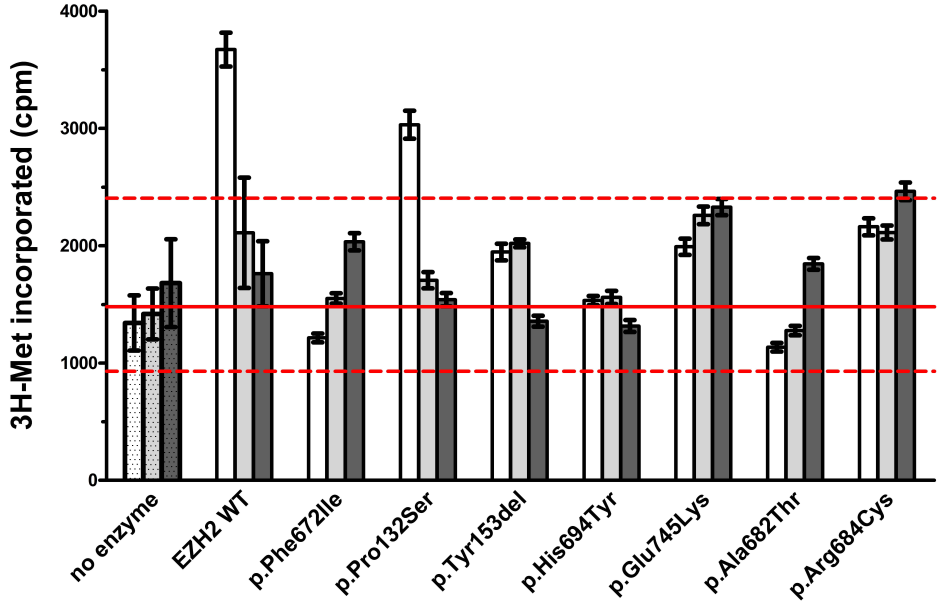



**Figure 3-4: Histone methyltransferase activity *in vitro* assay using differentially methylated substrates confirmed impaired activity, particularly with reduced ability for monomethylation of H3K27.**
Histone methyltransferase reactions were performed using 1 µM biotinylated peptide substrate (H3(21-44) mimicking the H3 tail) which had been either unmethylated (H3K27me0, open bars), monomethylated (H3K27me1, light gray bars) or dimethylated (H3K27me2, dark gray bars) at lysine residue 27. The reactions were incubated with 0.67 µM $^{3}$H-S-Adenosyl-methionine ($^{3}$H-SAM) and 250 ng of either wild-type (WT) or a mutant complex (or no enzyme controls). Histone methyltransferase activity was measured based on the incorporation of $^{3}$H-labeled methyl groups, represented in scintillation counts per minute (total counts). Error bars represent calculated standard deviations of two independent replicates for each mutant complex, and four replicates for WT. The red lines represent mean background (solid line), and minimal or maximum background (dotted lines) observed in this experiment. Further statistical data on the background is presented in Appendix I.4.

### 3.4.2 Methyltransferase activity of p.(Asp185His)

As described in Chapter 2, section 2.3.2.2.2, the p.(Asp185His) or D185H variant was present in a total of ten individuals who were referred for WS-like features and overgrowth. Two of these individuals also carried a pathogenic mutation in *EZH2* that could explain their overgrowth and dysmorphic features, but the phenotype of the remaining eight individuals cannot be caused exclusively by this common variant because there are far too many "healthy"

D185H carriers (~8% of the general population, as reported in various public databases; see Table 2-2).

The D185H variant has now been associated with three distinct phenotypes: it is present constitutionally in multiple healthy controls and in individuals with generalized overgrowth, and somatically in individuals who developed acute leukemias.[57,312] The lack of a specific phenotype-genotype association makes it challenging to determine the real effect of this variant. Importantly, recent genome-wide association studies (GWAS) for human height,[313–316] head circumference,[317,318] cognition,[319] and facial morphology [320–322] have not identified this variant as a contributing factor for any of these quantitative traits.[314,317] Similarly, studies investigating the D185H variant specifically as a potential risk allele for cancer did not find a significant association between carrier status and development of prostate or lung cancer,[323,324] with one study even suggesting a reduced risk for cancer (although this was a relatively small study with 523 cases and 523 matched controls, and thus had weak statistical power).[325] Together, this evidence supports the classification of the D185H variant as a common benign SNP (single nucleotide polymorphism) rather than "likely benign". Accordingly, this variant has recently been classified as benign in ClinVar (see Table 2-2 for reference), which also meets the criteria for variant interpretation established by the American College of Medical Genetics and Genomics (ACMG). In this context, the fact that our functional work suggests impaired histone methyltransferase activity for this mutant is notable. Based on careful repetition of our assays with different lot numbers of the wild type protein and under varied conditions, we do believe our *in vitro* results to be reproducible. Given the reproducibility of our assay in our hands and the use of a similar assay by multiple other groups, we do not believe the low methyltransferase activity exhibited by the D185H protein variant in this assay is indicative of experimental error. Rather, we believe that our results reflect the true activity of this variant under this select set of experimental conditions. Nevertheless, given the lack of a specific phenotype for D185H carriers, these *in vitro* results may not reflect its true activity *in vivo*.

On this basis, we must conclude that this particular *in vitro* assay cannot be used in isolation to assess the potential pathogenicity of novel *EZH2* variants. Instead, pathogenicity should be assessed based on the sum total of available evidence from family-specific co-segregation with the disease phenotype, population genetics and, where available, other orthogonal lines of functional evidence. The possibility of "pseudodeficiency alleles" is an uncommon but known

phenomenon whereby some protein variants manifest impaired activity by *in vitro* assays but have no demonstrable phenotypic effects *in vivo*.[326–328] Alternatively, our results could indicate a disruption in PRC2 assembly which was missed during quality control at BPS Biosciences, such that the functional impairment observed would be due to the complex "falling apart" rather than to true EZH2 impairment. In such case, the same possibility would have to be considered for all our mutant EZH2-PRC2 complexes.

To understand the discrepancy between the predicted (normal) activity of this common protein variant and its observed (deficient) activity more fully, we will require more definitive studies such as determination of the binding constant ($K_m$) for substrates, assays over the linear portion of the product vs. time curve, and careful sub-studies of the different enzymatic steps for a variety of different rare and common protein variants. In addition, we should carry out exome sequencing for the D185H carriers who did not also carry a pathogenic mutation in *EZH2* to look for other coding variants that may explain their overgrowth and dysmorphic features. The identification of other causes of disease in these individuals would support a definitive classification of the D185H variant as benign, and would help us rule out any residual doubt that the functional impairment observed *in vitro* actually reflects a real (but mild) functional impairment *in vivo*, possibly with reduced penetrance. Finally, further investigations regarding EZH2 function in these individuals will require the use of patient-derived samples, as described in the previous section.

### 3.4.3 Histone methyltransferase activity in this *in vitro* assay does not correlate with phenotypic severity

We chose to assay EZH2 protein variants that represented a wide variety of WS phenotypes. We had hypothesized that more severe clinical features of WS (such as cerebral migration defects or the development of malignancy) might be associated with mutations in specific protein domains that were in turn associated with more striking alterations of histone methyltransferase activity. However, we observed no clear correlation between these parameters. We also observed no correlation between clinical severity and profiles of substrate specificity (Figure 3-4). These results are not surprising considering that phenotypic presentation is variable between patients with the same *EZH2* mutation. Furthermore, our results are consistent with Guglielmelli *et al.*[305]

who observed no correlation between *EZH2* mutational status and hematologic or clinical parameters in patients with myelofibrosis.

This lack of phenotype/genotype correlation suggests that factors apart from EZH2's H3K27 methyltransferase function might explain the phenotypic differences observed in WS patients. Such factors may be stochastic, genetic (for example due to modifier genes), or biochemical, in relation to other histone modifications such as H3K4me3 and H3K27ac (see Chapter 1, section 1.2.1.1.2 and Figure 1-3). Activity of accessory proteins that are absent from our *in vitro* assay might also change the conformation of the PRC2 complex and influence its affinity to bind to H3K27, thus altering the resulting phenotype. Such factors include the presence of JARID2 or PHF1,[196,197,329] or proteins that modify EZH2 post-translationally such as AKT.[208] Alternatively, the activity of WS mutants on other histone substrates such as H1BK26 (lysine 26 on histone H1B),[201] or non-histone substrates, such as STAT3,[330] JARID2,[331] GATA4[332] and RORα,[333] may be a more important determinant of the ultimate phenotype of WS patients.

# Chapter 4: Identification of *EED* as a novel overgrowth gene via exome sequencing

## 4.1 Rationale

In 2009, Ng *et al*. demonstrated the successful application of exome sequencing for identifying disease causing variants of a Mendelian disease.[334] A year later, the same group reported on the first successful gene discovery using exome sequencing, when they identified the cause of Miller syndrome (OMIM #263750).[335] Since then, "whole" exome sequencing (WES) has become the standard technique for investigating the cause of rare disorders,[1,3,291,292,336,337] mainly because the vast majority of Mendelian disorders appear to be caused by mutations that affect protein function.[291] With a drastic decrease in costs, this method has now become a standard in clinical diagnostics also,[1,292,338–341] and the overall diagnostic success rate reported using WES (in patients for which chromosomal microarray had been deemed normal) is around 25%.[3,338–341] One of the clear limiting factors in the field of rare diseases is the small number of patients available for investigations within a single study, highlighting a need for international collaborations.[1,291] Detailed phenotyping and delineation of sub-groups of patients with significant overlap in clinical features have been shown to aid in gene discovery;[291,292,342] this was the strategy used in our study.

Further, mutations in chromatin regulators such as *NSD1* and *EZH2* have been shown to cause the "classical" overgrowth syndromes Sotos and Weaver,[56,57,128] and more recently, *de novo* mutations in *DNMT3A* (DNA methyltransferase 3A; OMIM *602769) and *SETD2* (SET-domain containing protein 2; OMIM *612778), two other chromatin regulators, were shown to cause distinct overgrowth syndromes using WES.[160,161] Based on this evidence, we hypothesized that rare *de novo* mutations in other chromatin regulators, and particularly other members of the Polycomb Repressive Complex 2 (PRC2), might explain unsolved cases of Weaver-like overgrowth. Due to the overwhelming success of using WES for novel gene discoveries in the field of rare diseases, this was the method selected for our investigations.

### 4.1.1 Next generation sequencing strategy

Probands for which *EZH2* and *NSD1* testing by Sanger sequencing were negative (48/66) were prioritized for exome sequencing, a next-generation sequencing tool that allows sequencing of the entire protein coding portion of the genome (~1-2% of the total genomic space) at once and in a cost-effective manner.[343] A total of ten patients were selected based on phenotypic severity and available funding at the time of prioritization. Ten unrelated singletons were sequenced, hoping that variants in the same gene would be detected in unrelated individuals with sufficient phenotypic overlap (based on our targeted patient recruitment). Exome sequencing data were also available for one more individual (case 92) that had been sequenced previously.

### 4.1.2 Prior candidates

Before carrying out exome sequencing, a list of gene candidates was compiled based on their known function. In particular, other members of PRC2, proteins interacting with PRC2, other SET-domain containing proteins and other chromatin regulators were selected; a compilation of the top candidates and the reasoning for picking each gene is available in Appendix J. This list provided a starting point for the analysis of the exome sequencing data while still allowing for discovery of causative variants in other genes.

## 4.2 Methods

### 4.2.1 Exome sequencing

Library preparation, exome sequencing and bioinformatics analysis were carried out at Canada's Michael Smith Genomes Sciences Centre (GSC). Exome libraries were constructed using the Agilent All Exon V5+UTR capture kit and sequenced on Illumina HiSeq 2500. Reads were aligned to the human genome (GRCh37-lite) using Burrows-Wheeler Aligner (BWA).[344] Variants were called using mpileup (SAMtools),[345] subsequently filtered with varFilter, and annotated exome variant calling and annotation were performed using an in-house pipeline that combines SnpEff,[346] Ensembl variant database, dbSNP,[347] NHLBI exome sequencing data (https://esp.gs.washington.edu/drupal/), COSMIC,[348] and an in-house human variation database,[349] which consisted of 1643 healthy individuals and 819 cancer tissues. Variants unique

to our patient (defined as not previously observed within the in-house human variation database) were prioritized on a Microsoft Excel spreadsheet.

### 4.2.2 Coverage check

During library preparation for exome sequencing, capture is not equal for all regions of the exome, and as such not every exon of every coding gene generates a sufficient number of alignable reads to accurately call the genotype at that locus.[340,343] This remains one of the biggest technical limitation of WES and is slowly being tackled with technical improvements.[3] To address any possible gaps in coverage, we manually checked the read alignment for all exons of the most common known overgrowth genes, as well as novel overgrowth genes recently discovered in cohorts similar to ours (see Chapter 1, section 1.1.3); this was done using the freely-available Genome Browse software from Golden Helix. No major gaps in coverage were detected (summary available in Appendix K). In theory, this coverage check could also pick up copy number changes (by comparing exome data generated in parallel on the same run on the Illumina HiSeq 2500), but no clear indication of CNVs at these loci was observed.

### 4.2.3 Criteria for prioritizing good candidates from exome data

Exome sequencing analysis was carried out as described above, with the assumption that we were searching for a *de novo* heterozygous dominant rare variant in each exome. For each individual, between 190 and 375 constitutional variants unique to the proband (as defined in section 4.2.1) were called. These variants were further prioritized based on the following criteria, in the order presented:

1- variants in known overgrowth genes and in the genes from the prior candidate list were interpreted first by looking at type of genetic alteration, population frequency, *in silico* prediction scores, and coverage at that specific locus;

2- synonymous variants were excluded, while missense, frameshift, truncating/nonsense and potential splice-site variants were considered;

3- all variants with a minor allele frequency (MAF) over 1% in dbSNP, ExAc or the Exome Variant Server were excluded (because these are rare syndromes with congenital features);

4- variants with a reported "rs number", meaning that they were described in the dbSNP database, were not automatically excluded, but rather each entry on dbSNP was manually checked to inform on actual MAF and previous classification (see section 4.5.1);

5- variants in genes with an associated OMIM number were considered if the known gene function and/or associated disease phenotype matched that of the individual;

6- variants in genes with no known associated human disease but a known function reported in scientific literature that was consistent with the phenotype were also considered.

### 4.2.4 Validation of candidate variants

Top candidate variants (prioritized using the criteria described above) were validated by Sanger sequencing in the proband. Primer design, PCR optimization, sequencing preparation and analysis were all carried out as previously described for *EZH2* and *NSD1* screening. When a variant was confirmed in the proband, parents and siblings were also tested (if available) to inform on inheritance.

## 4.3 Discovery of *EED* as a novel overgrowth gene

Here we describe two patients suspected clinically to have Weaver syndrome but whose features are caused by constitutional mutations in *EED*, another member of the Polycomb Repressive Complex 2 (PRC2). EED partners with EZH2 within PRC2, and this interaction is required for proper EZH2-mediated histone methyltransferase activity that maintains gene silencing (Figure 4-1a).[189,350] The two male patients described here show significant phenotypic overlap, presenting with overgrowth, facial dysmorphism and intellectual disability. They carry different mutations in *EED* that are both rare and *de novo* (Figure 4-1b). These patients represent the first two reports of overgrowth and characteristic dysmorphism associated with constitutional mutations in *EED*.

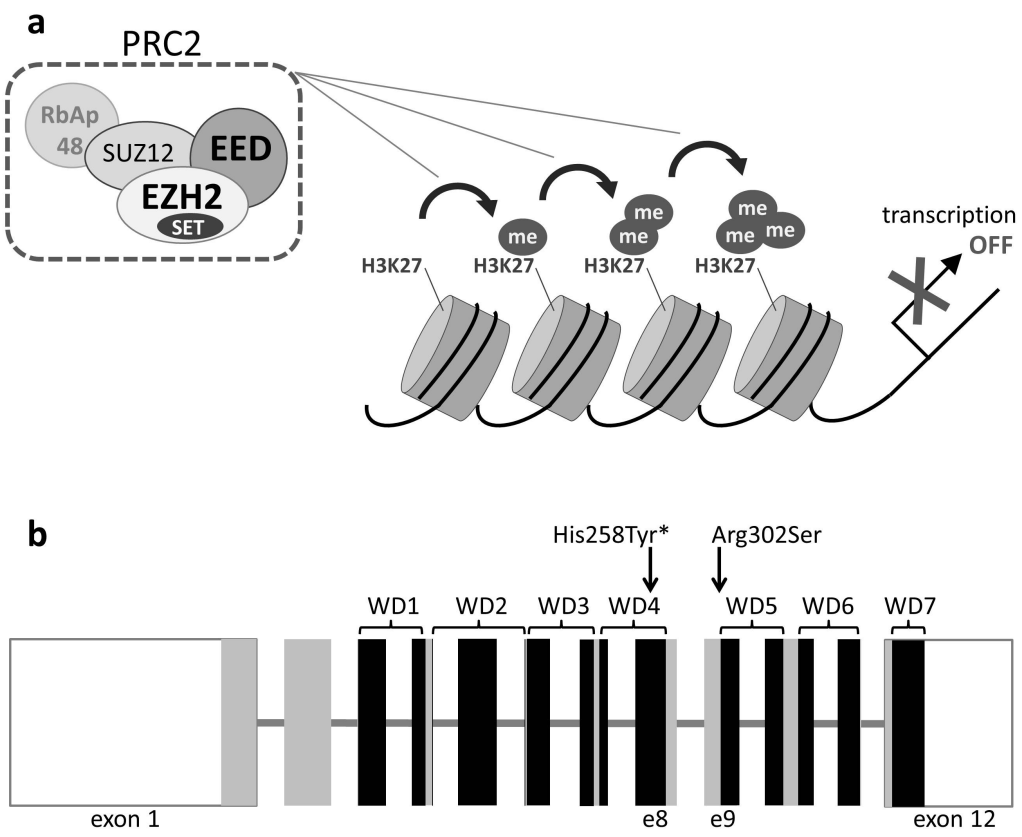



**Figure 4-1: Schematic of human EED and its role within the Polycomb Repressive Complex 2 (PRC2).**
(a) Schematic representation of PRC2. EED is required (along with EZH2 and SUZ12) for proper histone methyltransferase activity mediated by the SET domain of EZH2. Within PRC2, EZH2 can add up to 3 methyl groups (me) to lysine 27 on the histone 3 tail (H3K27). This is done in a sequential manner and shuts off transcription, leading to repression of gene expression. Disruption of EED is thought to disturb this PRC2-mediated histone methyltransferase activity. (b) Human EED is represented. Each rectangle represents one exon. Exon size is represented to scale, while intronic distances are not to scale. White (open) rectangles represent non-coding untranslated regions (UTRs) and grey rectangles represent coding exons (NM_003797.4). EED protein contains 441 amino acids (NP_003788.2) and seven WD repeats, represented here in black according to UniProt (O75530) coordinates. The two constitutional mutations in *EED* associated with overgrowth are shown (case 2 is indicated by an asterisk).

### 4.3.1 Clinical report of case 1

#### 4.3.1.1 Birth and childhood

Our proband was born full term to non-consanguineous Turkish parents. Birth weight was 4100 g (+1.09 S.D., standard deviations) and birth length 52 cm (+0.38 S.D.). Birth head circumference was not recorded. Dysmorphic features noted in childhood included macrocephaly, large bifrontal diameter, ocular hypertelorism and large ears (Figure 4-2a and b). Mild intellectual disability (IQ: 60) with speech delay (first words at age 2 years, first sentences at age 5 years) and poor fine motor skills were present. Brain MRI was normal. He developed epilepsy, with his first seizure occurring at 4.5 years. Electroencephalogram showed a generalized irregular wave pattern, and phenytoin and primidone reduced seizure frequency to 1-2 per year (with no seizures from age 27 onwards). He also developed moderate-to-severe myopia (prescription lenses -5.5 and -6.5 diopters). Echocardiogram found mild mitral regurgitation.

He developed tall stature in childhood: height was 87 cm (+4.2 S.D.) at 13 months, 138 cm (+5.1 S.D.) at 5 years 7 months and 158 cm (+4.6 S.D.) at 8 years 8 months (WHO growth charts). Blood hormone levels including thyroid-stimulating hormone (1.4 IU/ml), T4 (1.22 ng/ml), insulin-like growth factor 1 (201 ng/ml) and fasting glucose (98 mg/dl or 5.4 mmol/l) were all normal. He also had an umbilical hernia, and required surgical correction for cryptorchidism and a post-traumatic patellar dislocation at age 9 years. Following his knee surgery, circulatory failure led to amputation of his right leg over the knee at age 11 years.

X-rays at age 4 years revealed dense physeal bands within the proximal femurs, and at 14.5 years bone age was consistent with 15 years. Additional X-rays at various ages revealed significant scoliosis, abnormal flaring of the distal clavicles, distal ribs, and metaphyses of the distal radius, distal ulna, distal femur and proximal tibia. The metaphyses were also abnormally lucent. He had flattened glenoid fossae and humeral heads, as well as a flattened left acetabulum and femoral head.

#### 4.3.1.2 Adult presentation

At the most recent examination at age 27, his final height was 190 cm (+1.85 S.D.) and head circumference 59 cm (+1.46 S.D.). He has since undergone bilateral cataract surgery (age 30 years 10 months). Photographs of the patient confirm dysmorphic features including

hypertelorism (adult interpupillary distance 8 cm), downslanting palpebral fissures and retrognathia with a prominent crease between the lower lip and the chin (Figure 4-2a-h, Table 4-3). He also has kyphoscoliosis, widely spaced nipples and several pigmented nevi (Figure 4-2e,f), as well as large hands with camptodactyly (Figure 4-2h). To date, he has had no malignant or premalignant phenotypes and his blood cell counts remain normal (red cells 4.9 x $10^{12}$ per litre, white cells 6.5 x $10^{9}$ per litre and platelets 250 x $10^{9}$ per litre). He has a younger sister who does not have any overgrowth or dysmorphic features.

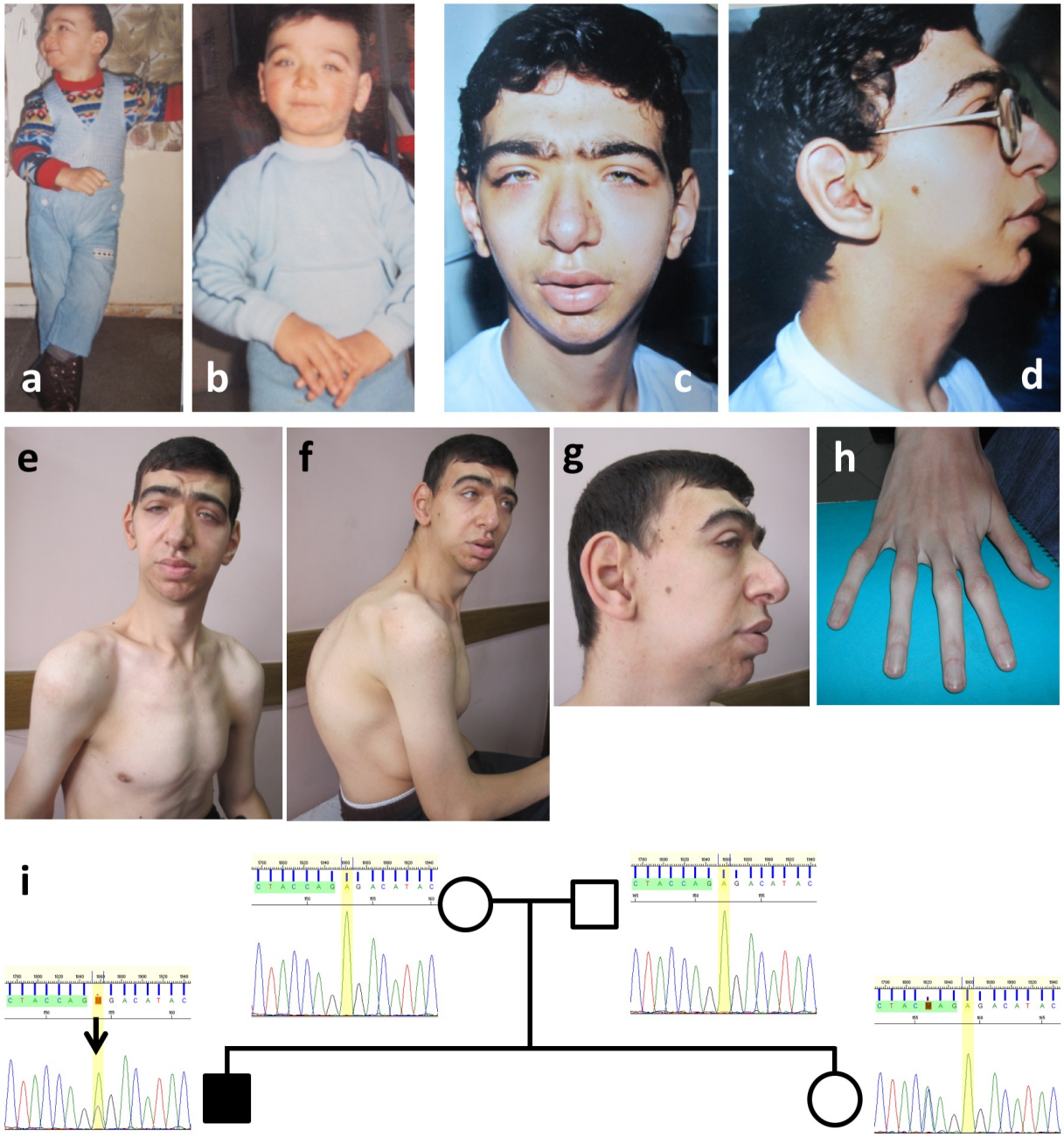



**Figure 4-2: Characterization of case 1, the first described patient with a constitutional *EED* mutation.**
(a-h) Photographs of the proband show the features described in the paper at various ages: 3 years (a, b), 14 years (c, d) and 30 years (e-g); proband's right hand is shown at 27 years of age (h). (i) Pedigree with Sanger confirmation that the patient carries a *de novo* c.1372A>C (p.Arg302Ser) mutation in EED. This mutation is absent in the sister who carries a variant in the adjacent residue; this variant is synonymous (p.Thr301=) and thus consistent with her lack of overgrowth and dysmorphic features. Sanger traces were analyzed using Sequence Scanner v1.0.

### 4.3.2 Clinical report of case 2

#### 4.3.2.1 Birth and early years

Our proband required forceps-assisted delivery after an uncomplicated term pregnancy (42 weeks by dates). His parents are non-consanguineous Caucasians and have younger healthy twin sons; there is no family history of overgrowth. The father was 36 and the mother 32 years at conception. Ultrasound at the beginning of the third trimester identified macrosomia. Birth weight was 4366 g, length 54.6 cm and head circumference 37.2 cm (Table 4-3). Apgar scores were $3^{(1min)}$ and $4^{(5min)}$. He had respiratory distress and mild jaundice. An umbilical hernia developed one week after birth. Developmental delay was apparent early: he could only say one word by 14 months and 2 words by 19 months. At 17 months he could feed himself and hold himself up; he crawled a few weeks later. When standing, his legs and Achilles tendons were stiff and he stood primarily on his toes; physiotherapy improved his range of movement. Karyotype was normal (46, XY), and he was referred to Medical Genetics where his delayed motor skills, cognitive difficulties, large size and dysmorphic facies (Figure 4-3a-c and f), suggested Weaver syndrome. At 20 months, he took four steps alone. At 22 months, he got casts for heel cord lengthening and said his 3$^{rd}$ word. At 24 months he could walk unassisted. At 26 months, he started learning sign language, and at 30 months he could say 3 more words. He could also go up and down stairs unassisted.

#### 4.3.2.2 Childhood

At 5 years, some asymmetry of the skull was noted and X-rays revealed a bone age of 8 years. At 6 years 1 month, his verbal scores remained equivalent to age 2.5 years. Insulin-like growth factor 1 levels were normal at age 6 years 5 months. He could ride a bike at age 7 years. By the time he reached third grade (around age 8 years), his speech was much improved and he interacted socially with his peers. At 8 years 8 months (Figure 4-3g), his overall IQ was 52 (verbal IQ: 64, performance IQ: 47; Wechsler Intelligence Scale for Children - III): he could not read or recognize numbers but could count objects, and specific weaknesses were noted in problem solving and memory (except for visual memory). Caregivers found him to be socially interactive and very personable. Slowness with upper extremity motor skills (attributed to dyspraxia) made it hard for him to write, though he could copy drawings and write his own name. Hypotonia also contributed to coordination and balance difficulties, with marked

pronation of the feet and bent knees, in turn affecting the posture of his hips and back. Other biomechanical variations (rigidity of the first metatarsal ray bilaterally and hypermobility of the fourth and fifth rays of the lower limbs, and limited midfoot locking bilaterally) also affected his joint stability.

At 8 years 10 months, reduced range of motion in large joints was observed. Physical therapy (movement and gait training) continued. At 9 years 4 months, dental examination revealed excess overbite and overjet, deep anterior bite with impingement, and a vertical facial growth pattern with moderate mandibular retrognathia. Treatment included headgear and retainers for several years. By the fifth grade (approximately 10 years of age), his speech had improved and was rated as clear to unfamiliar individuals approximately 80% of the time. At 10 years 4 months, X-rays of the spine showed thoracolumbar scoliosis (18°), and X-rays of the hand and wrist again showed advanced bone age (consistent with 12 years 6 months) and moderate osteopenia. Abdominal ultrasound was normal. His scores on the Vineland Adaptive Behaviour Scales (measured at 12 years 3 months) assessed his communication and daily living skills to be at a 6+ year level, gross motor and fine motor skills at a 4-5 year level, adaptive skills at a 6+ year level, and social skills at 7+ year level.

He had chronic constipation requiring Citrucel and high fibre cereals. Notable dysmorphisms included large hands with long slender fingers (middle finger length 9.5 cm, total hand length 20 cm) and large slender feet (total length 28 cm), macrocephaly, almond-shaped palpebral fissures, and bifid uvula. His rate of growth had slowed, with measurements on the 75$^{th}$ percentile (height 159.7 cm, weight 46 kg and head circumference 57.2 cm) (Appendix L). Weaver syndrome remained the most plausible recognized diagnosis for his phenotype (Figure 4-3h). At 14 years 2 months, reading was equivalent to a first grade level.

#### 4.3.2.3 Trauma, surgery and recovery

At age 15 years 6 months, the proband suffered neck trauma: he did a forward roll in gymnastics class and had immediate onset of gait ataxia, with numbness and weakness of his extremities. MRI revealed spinal cord compression at the occipitocervical junction. Flexion extension radiographs revealed significant C1-C2 instability, a substantially increased atlanto-dens interval, and assimilation of the atlas. Surgical treatment included suboccipital craniectomy, C1 and C2 laminectomies and lysis of dural band adhesions. Follow-up examination at two and

half months after surgery showed functional recovery; physical therapy eventually corrected his residual cervical misalignment.

#### 4.3.2.4 Adult years

Brain MRI at 22 years of age showed no evidence of impingement on the medulla or proximal cervical cord following surgical fusion and posterior decompression. At the latest examination (age 30 years 4 months), he was doing well and communicated verbally equivalent to a first grade level. He enjoyed meeting new people and had a good sense of humour. His adult height was 191 cm (>97$^{th}$ percentile), weight 93.4 kg (90-95$^{th}$ percentile) and head circumference 61 cm (>90$^{th}$ percentile). His stance remained slightly forward, with hips and knees flexed. Range of movement of certain joints was restricted: he was unable to reach his arms up over his head or to bend down to tie his shoes, with his heel cords remaining very tight. His fingernails and toenails were very fragile. Further features are shown in Figure 4-3i-l.

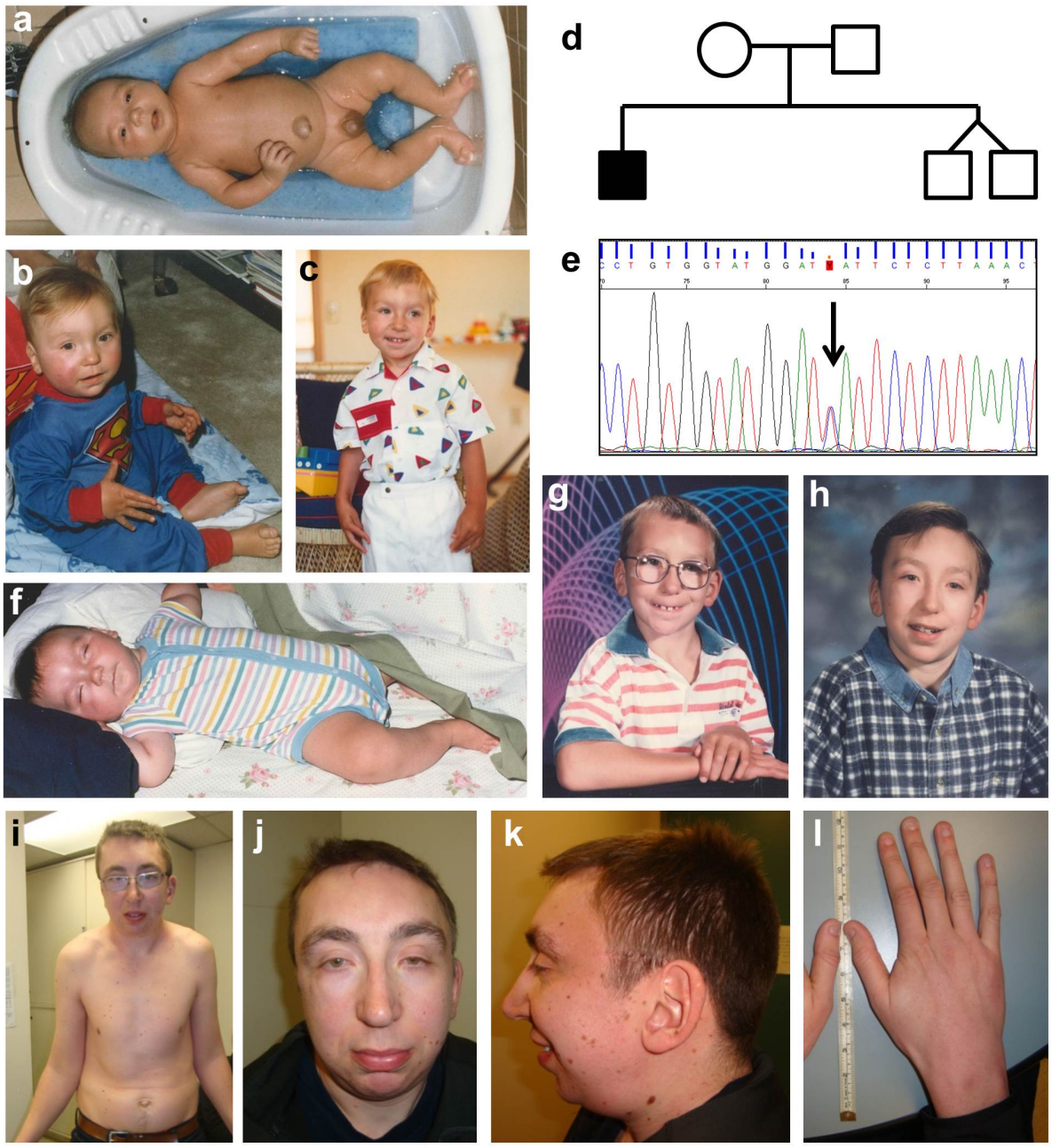



**Figure 4-3: Characterization of case 2, the second described patient with *EED*-related overgrowth.**
(a-c) Photographs of the proband at age 6 weeks (a), 6 months (b) and 2 and a half years (c) show that early features were consistent with Weaver syndrome. Note the rounded face, macrocephaly, retrognathia with "stuck-on" chin, long and slender nose, and large low-set ears. (d) Pedigree of the family showing that the proband is the only affected individual. (e) Sanger sequencing identified a *de novo* c.1238C>T (p.His258Tyr) mutation in *EED*, exclusive to the affected proband. (f-l) Photographs of the proband at various ages illustrate an evolving phenotype. Typical features of Weaver syndrome observed in early life (f: 7 months of age) remained apparent through

childhood (g: 8 years, h: 12 years). Recent photographs at age 30 years and 4 months (i-l) show dysmorphic features in adulthood. The main features include deep-set eyes, large low-set ears, prominent nasal root and nasal bridge with bulbous nasal tip, and retrognathia with a prominent crease between the lower lip and the chin (j,k). His thorax is narrow and pivoted forward on his hips slightly, and slight kyphosis is apparent (i). He also has numerous pigmented nevi across his chest (i) and face (k). The proband's right hand shown at 30 years and 4 months (l) is unusually large, measuring 23.5 cm. The wrist is broad, fingers are long and slender with long phalanges and very thick skin over the knuckles, and fingernails are fragile and paper-thin.

### 4.3.3 Sequencing and results

#### 4.3.3.1 Exome sequencing in case 1

Through exome sequencing (see section 4.2.1), we identified in case 1 a novel c.1372A>C missense variant in *EED* (NM_003797.3), confirmed to be *de novo* and absent in the healthy sister by Sanger sequencing (Figure 4-2i). This variant is predicted to convert arginine residue 302 to serine (p.Arg302Ser), and is scored as damaging by Polyphen although not by SIFT (Table 4-1).

We also found and Sanger-validated two novel variants in *MEGF8* (NM_001410.2): c.4517C>A inherited from the mother and c.3567G>A from the father, both absent in the sister (Figure 4-4). Although this gene has been associated with an autosomal recessive subtype of Carpenter syndrome (OMIM #614976),[351] which does feature some of the phenotypic traits observed in our patient (e.g. intellectual disability, cryptorchidism and camptodactyly), our proband had none of the cardinal manifestations of this latter syndrome (lacking polydactyly, heterotaxia or obvious craniosynostosis), and both *MEGF8* variants were predicted to be benign (Table 4-1). We found no other coding variants that could plausibly explain the observed features; therefore, we believe the *de novo EED* variant is the most likely cause of this patient's excessive height in childhood (max +5.1 S.D.).

| Gene | Chr | Position (build 37) | Mutation | Amino acid change | Inheritance | Polyphen/ SIFT prediction | Public database reference* | In-house databases ** |
|---|---|---|---|---|---|---|---|---|
| *EED* | 11 | 85979543 | c.1372A>C | p.Arg302Ser | *de novo* | probably damaging/ tolerated | not described | not described |
| *MEGF8* | 19 | 42856391 | c.3567G>A | p.Gly978Ser | paternal | benign/ tolerated | not described | not described |
| *MEGF8* | 19 | 42858811 | c.4517C>A | p.Ser1294Arg | maternal | benign/ tolerated | not described | not described |

**Table 4-1: Summary of candidate variants in this proband.**

* Public databases checked include: dbNSP, COSMIC, ExAc, Exome Variant Server, Ensembl variant database, NHLBI exome sequencing data, and LOVD.

** In-house databases refer to Canada's Michael Smith Genome Sciences Centre's human variation database and a Turkish-specific population database consisting of 587 exomes from the TÜBİTAK Advanced Genomics and Bioinformatics Research infrastructure (http://www.igbam.bilgem.tubitak.gov.tr/en/index.html).

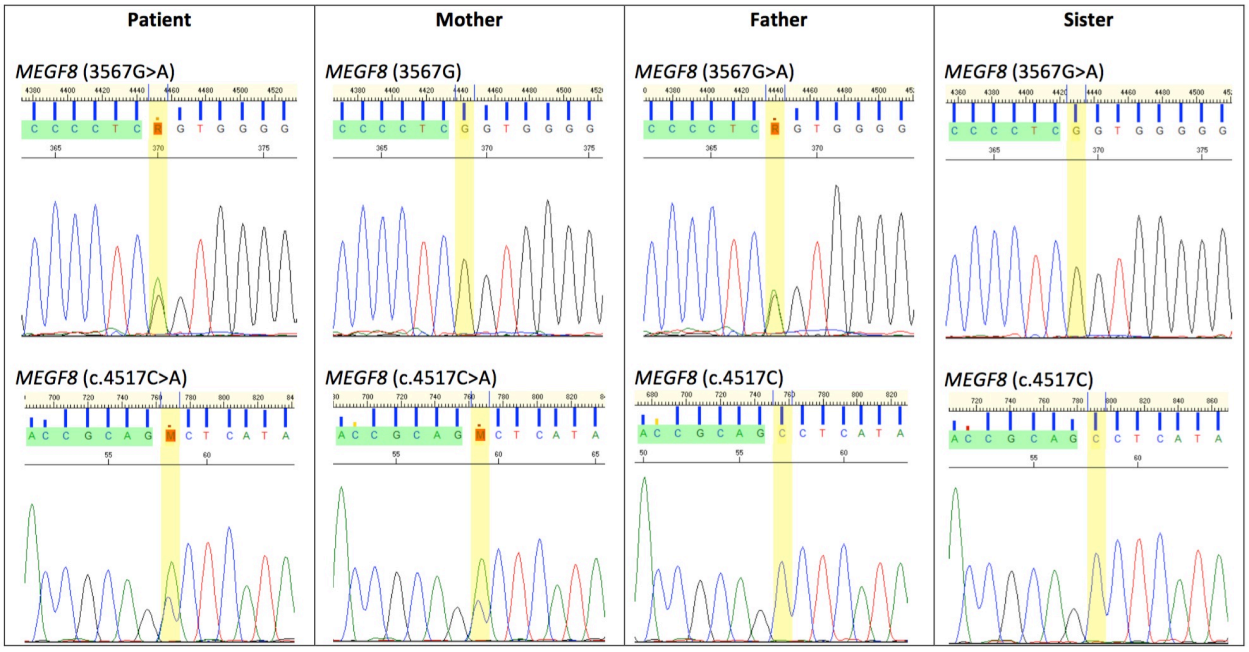



**Figure 4-4: Sanger validation of the *MEGF8* variants in the quartet.**

Sanger confirmation that case 1 is a compound heterozygote for two rare variants in *MEGF8*. Both variants are absent in the healthy sister. Sanger traces were analyzed using Sequence Scanner v1.0.

#### 4.3.3.2 Sanger sequencing in case 2

Following our discovery that a constitutional mutation in *EED* caused Weaver-like features in case 1, we screened all coding exons of *EED* (Figure 4-1b) by Sanger Sequencing (primers designed as before and described in Appendix C.3, PCR conditions described in D.4) in a total of 21 individuals from our Weaver-like cohort that had tested negative for rare variants in *EZH2* and *NSD1*. In this patient only (case 2), we identified a c.772 C>T missense variant (c.1238 C>T in the full-length mRNA sequence NM_003797.3), that was absent from his parents and siblings (Figure 4-3d,e). This variant, predicted to convert histidine residue 258 to tyrosine (p.His258Tyr), is not reported in the dbSNP or COSMIC databases, nor in ExAc or the Exome Variant Server, and was predicted damaging by both PROVEAN and SIFT. It is thus both a novel variant and a *de novo* mutation. As such, case 2 represents the second report of a rare *de novo* constitutional mutation in *EED* associated with overgrowth, intellectual disability and dysmorphic features.

#### 4.3.3.3 Additional results

Surprisingly, when carrying out validation of the *EED* (p.Arg302Ser) variant in the family of case 1, we identified a different *de novo* variant in *EED* in the proband's sister (see Figure 4-2i). This alteration is located just 3 bp upstream of the variant observed in our proband, thus within the adjacent codon, and is predicted to be synonymous (p.Thr301=). This variant has been observed in one individual from the ExAc population, with a corresponding entry in dbSNP (rs775659115), and is described in COSMIC (COSM467512) as being present in homozygous state in a renal cell carcinoma primary tumour sample.

Given that this is a synonymous variant, we do not expect it to have an effect on protein function, which is consistent with the lack of overgrowth and dysmorphic features in the sister. However, what is surprising, is that we would find two different rare *de novo* variants in *EED*, so close to each other, and within one family (family relationships between samples have been confirmed, as described earlier). This is even more surprising considering that alterations in *EED* are altogether very rare (for reference, there are only 129 SNVs reported in the 60,706 unrelated individuals from ExAc). These *de novo* variants are more likely to have arisen in the father's sperm, knowing that the rate of *de novo* mutations increases with father's age (father's age at

conception was 34 for the proband and 54 for the sister).[352,353] To address this hypothesis, we would need to establish parental haplotypes using surrounding SNP markers.

### 4.3.4 Discussion supporting *EED* as a novel overgrowth gene

Given the known effects on growth of coding mutations in *EZH2*, we posited a high prior probability that functional mutations in other members of the PRC2 complex, such as *EED*, would likely cause overgrowth. The evidence presented below supports this hypothesis.

#### 4.3.4.1 Conservation across species

EED is highly conserved in mammals with 100% amino-acid identity between mouse Eed and human EED, despite significant coding nucleotide differences between the *Eed* and *EED* genes.[354] This form of conservation suggests that disruption of any given residue is likely to affect protein function.

The fact that common amino acid polymorphisms in *EED* have not been observed across multiple healthy controls also suggests that sequence variation at the protein level is not likely to be compatible with good health, at least at the whole organism level. This is consistent with evidence derived from the ExAc study,[269] which shows that mutations in *EED* are rare, and predicts that *EED* is intolerant to missense mutations (z-score of 2.69, with 73 variants observed vs. 137.5 expected) and highly intolerant to loss-of-function (pLI of 1.00, with only one stop gained variant described in the last amino acid).

#### 4.3.4.2 Functional hypothesis

EED is required for proper methyltransferase activity of EZH2.[189] It contains seven WD40 domains (see Figure 4-1b),[203,355] which are functionally necessary to mediate interactions with other proteins, including EZH2.[189,350,356,357] Both mutations identified here (encoding p.Arg302Ser and p.His258Tyr) are located in highly conserved regions; [354,356] p.(His258Tyr) falls within the fourth WD domain,[355] and p.(Arg302Ser) localizes to the boundary of the fifth WD domain (see Figure 4-1b).[357] Sewalt *et al.* showed that all WD domains in human EED must remain intact for proper binding and interaction with EZH2,[350] and Montgomery *et al*. showed that deletion of individual WD domains disrupted PRC2-mediated H3K27 methylation in mouse embryonic stem cells.[357]

Independent lines of evidence also support functional importance of these residues. For example, a somatic mutation located one amino acid away from that mutated in case 2 (p.Ser259Phe) has been described in acute lymphoblastic leukemia,[233] and two loss-of-function mutations (*esc*$^{9}$ and *esc*$^{1}$) have been identified at nearby residues in the *Drosophila* orthologue (corresponding to Met256Lys and Leu260Arg in human EED).[358,359] Moreover, *in vitro* assays of the *Drosophila esc*$^{9}$ variant showed that the mutated protein did not bind E(Z) as efficiently as wild-type,[356] and Ketel *et al.* showed that complexes harbouring the *esc*$^{9}$ mutation have reduced histone methyltransferase activity.[360] Consistent with these observations, residue Met256 has been shown to be located on the outer surface of EED, at a prime position to interact with partner proteins.[355,356]

Similarly, the region surrounding residue 302 (mutated in case 1) appears to play an important role in binding directly to H3K27 through a small pocket on the surface of EED formed by hydrophobic residues including Cys324, Tyr364 and nearby residue Tyr308 (which in turn promotes further recruitment of PRC2 as well as of PRC1 to maintain nearby gene repression, as described in Chapter 1).[203,361] Interestingly, a mutation at Tyr358, six amino acids away from Tyr364 (the same distance as Arg302 to Tyr308), reduced EED binding to histone peptides by twofold.[203] Other lines of evidence supporting the hypothesis that these mutations disrupt EED function are summarized in Table 4-2 below.

| Organism | Gene | Gene disruption | Observations | Database entry | Reference |
|---|---|---|---|---|---|
| Human | *EED* | p.(Arg302Gly) (somatic) | Myeloproliferative neoplasm (a type of tumour that has recurrent mutations in other overgrowth genes like *EZH2* and *DNMT3A*) | COSMIC: COSM3720451 | [362] |
| Human | *EED* | Copy number loss 11q14.1-22.3 | Developmental delay and/or other significant developmental or morphological phenotypes | ClinVar: SCV000080064; dbVar: nsv531427 | [363] |
| Human | *EED* | Copy number loss 11q14.1-22.2 | Abnormal facial shape, Muscular hypotonia | ClinVar: SCV000080065; dbVar: nsv531428 | [363] |
| Human & Mouse | *EED* | Null allele (c.T1040C) and hypomorphic allele (c.T1031A) in the same WD domain | Mutations in construct contacting human EED disrupt direct interaction with mouse Ezh2 *in vitro* | non applicable | [364] |
| Mouse | *eed* | Hypomorphic allele *l7Rn5*$^{1989SB}$ (c.T1031A, p.N193I) in second WD domain | Heterozygous mutant skeletons display intermediate phenotype: non-lethal posterior transformations along the anterior-posterior axis. | non applicable | [222] |

| Organism | Gene | Gene disruption | Observations | Database entry | Reference |
|---|---|---|---|---|---|
| Drosophila | *esc* | Site-directed mutants RDE216AAA, GG210AA and double mutant RED216AAA DFST278AFAA | Mutant esc proteins that affect residues on the surface of the protein and that showed reduced binding to e(z) *in vitro* are capable of associating in complex with e(z) but have impaired function *in vivo*. | non applicable | [365] |
| Drosophila | *esc* | Various mutants within WD domains including $esc^5$ (p.Q171STOP), $esc^9$ (p.M236K) and $esc^2$ (frameshift insertion and deletion at $V_{404}$) | De-repression of homeotic genes resulting in transformation of body segments, for all mutants, despite some alterations being in "non-conserved" residues, suggesting that individual WD repeats are all important and may have specialized roles. | non applicable | [358] |

**Table 4-2: Genotype/Phenotype correlations of the *EED* gene (11q14.2) and orthologues.**

Together, these data suggest that coding mutations in *EED* (particularly within WD domains) are likely to affect protein function and, in turn, disrupt PRC2-mediated H3K27 methylation. Thus, constitutional *EED* mutations could lead to overgrowth via similar molecular mechanisms to constitutional *EZH2* mutations, which appear to reduce H3K27 methylation,[59] potentially causing derepression of Hox genes during embryonic development.[189,358,359,366] The pathophysiological mechanism is not yet understood and will require further investigation (see Chapter 5, section 5.4.2).

#### 4.3.4.3 Defining a new overgrowth syndrome

Based on the evidence discussed above and the fact that we have identified two different rare and *de novo* mutations in *EED* that are associated with overgrowth, intellectual disability and characteristic dysmorphism, we conclude that these are truly pathogenic variants and that we have successfully identified a new overgrowth gene. Both of our patients presented with Weaver-like features at an early age, including a rounded face with a "stuck-on" chin, overgrowth and intellectual disability (Table 4-3). It has been well established that many features attenuate with age in Weaver syndrome patients,[58] whereas these two individuals with *EED* mutations remain very dysmorphic in adulthood and have more severe skeletal perturbations, as well as restricted joint movement and unusually large hands. *EED*-associated overgrowth does not consistently

remain above the 95th percentile for stature and weight throughout childhood and adolescence (see Appendix L for growth curves of case 2), though adult height of both patients was above the 95th percentile (Table 4-3). Predisposition to haematological and other malignancies is known to occur in Weaver syndrome;[58,59,61] to date, neither patient with a *de novo EED* mutation has developed neoplasia, but the large number of cutaneous nevi in case 2 suggests the possibility of precancerous lesions. We will need to identify additional patients (particularly older adults) and follow them longitudinally to get better data on constitutional cancer predisposition, if any, conferred by mutations in *EED*. Additional cases will also allow us to characterize the full phenotypic spectrum of *EED*-associated overgrowth and to conclusively determine whether this is truly a distinct syndrome or simply a variation of Weaver syndrome (see Chapter 5, section 5.2 for further discussion).

| **Characteristics** | ***EED* case 1 (section 4.3.1)** | ***EED* case 2 (section 4.3.2)** |
|---|---|---|
| *Reference* | [367] | [368] |
| *De novo* mutation in *EED* | p.Arg302Ser | p.His258Tyr |
| **Growth features** | | |
| Gestational age at delivery (weeks) | 40 | 42 |
| Birth weight (kg) | 4.1 (+1.09 S.D., 86th %ile) | 4.37 (+0.68 S.D. for 42 weeks, 75th %ile; or +1.62 S.D. for a normal 40-week term, 95th %ile) |
| Birth length (cm) | 52 (+0.38 S.D., 65th %ile) | 54.6 (+0.77 S.D. for 42 weeks, 78th %ile; or +1.53 S.D. for a normal 40-week term, 94th %ile) |
| Birth head circumference (cm) | NK | 37.2 (+0.84 S.D. for 42 weeks, 80th %ile; or +1.52 S.D. for a normal 40-week term, 94th %ile) |
| Recent weight (kg) [age] | 85 [27y] (BMI 23.5 kg/m$^2$) * | 93.4 [30y4m] (BMI 25.6 kg/m$^2$) |
| Recent height (cm) [age] | 190 [27y] (+1.85 S.D.) | 191 [30y4m] (+2 S.D.) |
| Tall stature, maximum S.D. reached | +5.1 | +3.24 |
| Excessive growth of prenatal onset | - | + |
| Excessive growth of postnatal onset | + | - |
| Accelerated osseous maturation | +/- | + |
| **Neurological features** | | |
| Hypertonia | + | + |
| Hypotonia | - | + |
| Hoarse low-pitched cry | + | NK |
| Intellectual disability [IQ] | + [60] | + [52] |
| Excessive appetite | - | - |
| Ventriculomegaly | - | - |

| **Characteristics** | ***EED* case 1 (section 4.3.1)** | ***EED* case 2 (section 4.3.2)** |
|---|---|---|
| Delayed myelination | - | - |
| Cerebellar hypoplasia (mild) | - | - |
| Seizures [age of onset] | + [4.5y] | - |
| Polymicrogyria | - | - |
| Developmental delay | + | + |
| Poor fine motor coordination | + | + |
| Poor balance/gravitational insecurity | + | + |
| **Craniofacial** | | |
| Macrocephaly | + | + |
| Prominent forehead | - | + |
| Large bifrontal diameter | + | - |
| Flat occiput | + | - |
| Large ears | + | + |
| Low-set ears | - | + |
| Hearing loss | - | - |
| Ocular hypertelorism | + | - |
| Myopia | + | + |
| Strabismus | - | - |
| Cataracts | + | - |
| Down slanted palpebral fissures | + | - |
| Almond-shaped palpebral fissures | - | + |
| Full/ thick eyebrows | + | + |
| Long philtrum | - | - |
| Prominent nasal root and nasal bridge with bulbous nasal tip | + | + |
| Retro/Micrognathia | + | + |
| Prominent crease between lower lip and chin | + | + |
| “Stuck-on” chin in childhood | + | + |
| High palate | NK | + |
| Bifid uvula | NK | + |
| **Cardiovascular** | | |
| Patent ductus arteriosus | - | - |
| Ventricular or atrial septal defect | - | - |
| Mitral regurgitation | + (mild) | - |
| **Thorax** | | |
| Scoliosis | + | + (mild) |
| Kyphosis | + | - |
| Pectus carinatum or excavatum | - | - |
| Narrow thorax | NK | + |
| **Limbs** | | |
| Limited elbow and knee extension in early life | + | + |
| Limited elbow and knee extension after puberty | + | + |
| Metaphyseal Flaring | + | + |
| Flaring of distal clavicles and ribs | + | - |
| Flattening of ball-and-socket joints | + | - |
| Physeal Bands in Proximal Femurs | + | - |
| **Hands** | | |
| Large hands | + | + |
| Prominent digit pads | + | - |
| Single transverse palmar crease | - | + |
| Camptodactyly | + | - |

| Characteristics | *EED* case 1 (section 4.3.1) | *EED* case 2 (section 4.3.2) |
|---|---|---|
| Arachnodactyly | + | + |
| Broad thumbs | + | - |
| Thin, deep-set nails | + | + |
| **Feet** | | |
| Large feet | NK | + |
| Clinodactyly, toes | - | + |
| Talipes equinovarus | - | - |
| Short fourth metatarsals | - | - |
| Hind foot valgus | - | + |
| **Skin** | | |
| Excessive doughy/loose skin | - | - |
| Hypoplastic/supernumerary nipples | - | - |
| Widely-spaced nipples | + | + |
| Thin hair | - | - |
| Increased number of pigmented nevi | + | + |
| **Connective tissue** | | |
| Umbilical hernia | + | + |
| Inguinal hernia | - | - |
| Diastasis recti | + | - |
| Cryptorchidism | + | - |
| **Endocrine** | | |
| Hypothyroidism [age of onset] | - | - |
| Growth hormone deficiency [age of onset] | - | - |
| **Neoplasia** | | |
| Neuroblastoma, Leukemia or Lymphoma | - | - |

**Table 4-3: Detailed phenotypic comparison between the two individuals with constitutional mutations in *EED* associated with overgrowth.**

Key: + = present; - = assessed and found to be absent; +/-= present in early years then resolved; NK = not known; y = years; m = months; S.D.= standard deviations; %ile = percentile.

* Patient had had a leg amputation secondary to surgical complications at age 11 years.

## 4.4 Candidates identified and validated in the other completed exomes

The top candidate variants in each exome were selected for Sanger validation based on the criteria described in section 4.2.3. The list of all candidate variants validated by Sanger (and accompanying information such as bioinformatic prediction scores of pathogenicity) is available in Appendix M; the most relevant results are described below.

### 4.4.1 Variants in genes previously associated with overgrowth

One patient (case 37) was found to have a p.(Ala571Pro) missense variant in *DNMT3A* (NP_072046.2), confirmed *de novo* via Sanger sequencing. This variant is not reported in dbSNP, ExAc, ClinVar, DECIPHER or the Exome Variant Server. It is predicted damaging/deleterious by SIFT and Polyphen, but neutral by PROVEAN. Two different variants affecting the same amino acid have been reported in COSMIC (p.(Ala571fs*80) and p.(Ala571Ser)), suggesting possible pathogenicity for variants affecting this residue (although these are somatic mutations, not constitutional). Importantly, *DNMT3A* was recently described as an overgrowth gene.[160] In the original report, the authors stated that only variants disrupting known protein domains were associated with disease, and a recent report from an independent group (available online in June 2016) supported this claim,[369] even though the protein structure presented in the two studies varied slightly illustrating the need for standardized nomenclature across studies looking at the same gene/protein. The p.(Ala571Pro) variant identified in case 37 is located within an ADD zinc finger domain, and the closest variant reported in a patient with overgrowth is p.(Cys549Arg).[160] Unfortunately, the phenotypic information reported in these two studies is limited, not allowing us to determine whether there is sufficient overlap with the clinical presentation of our patient. Therefore, this variant remains classified as a VOUS, and further studies (or discovery of the same variant in another affected individual) will be required to determine whether the *DNMT3A* variant identified in case 37 is causative of his overgrowth phenotype.

Another patient (case 29) was found to have a missense variant in *SETD2*, a gene also recently linked to overgrowth.[161] This variant, p.(Thr159Ala), is now reported in dbSNP (rs369333306) as it was observed once within the ExAc population (of 60,706 individuals), and it is also reported in the Exome Variant Server with a frequency of 0.02% overall (but 0% in the European American population). It is predicted benign by Polyphen, SIFT and PROVEAN. Upon validation, this variant was found to be inherited from the father; he does not appear to show features of generalized overgrowth or dysmorphism, and overall facial features do not resemble those observed in the proband. However, information on the father's phenotype in childhood was not available, and many features of overgrowth syndromes are known to diminish with age. Further, there are only four reported individuals with constitutional mutations in *SETD2*, only one of these being missense, and the clinical presentation is extremely

variable.[161,162,369] Therefore, although this variant seems unlikely to be causative, current evidence is insufficient to definitively rule out the pathogenicity of the *SETD2* variant identified in case 29.

### 4.4.2 Candidate variant in a potentially novel overgrowth gene

An interesting new candidate gene for overgrowth was identified in a third patient (case 92). We detected a novel missense variant in *CHD3* (chromodomain helicase DNA-binding protein 3; OMIM *602120), also known as *Mi2-ALPHA*. This variant, p.(Arg985Gln), was confirmed to be *de novo* by Sanger sequencing and is predicted deleterious/damaging by PROVEAN, SIFT and Polyphen. Further, ExAc predicts this gene to be intolerant to loss-of-function and highly intolerant to missense mutations (z-score of 7.15).[269]

CHD3 is a member of the evolutionarily conserved CHD family of enzymes, belonging to the SNF2 superfamily of ATP-dependent chromatin remodelers.[370,371] Like all CHD proteins, CHD3 contains two chromodomains (chromatin organization modifiers) and a SWI/SNF2-like ATPase domain, plus an additional two PHD zinc-finger domains.[371] The variant identified in our patient is located in the SWI/SNF2-like domain, essential for enzymatic function.[372] CHD3 (Mi-2α) and/or CHD4 (Mi-2β) act as core subunits of a multiprotein complex called NuRD (nucleosome remodeling and histone deacetylase), or Mi-2/NuRD.[373,374] Within NuRD, CHD3/4 mediate the ATPase-dependent remodelling of nucleosomes, which in turn enhances the histone deacetylase function catalyzed by partner HDACs.[373,374] Coupling of these two independent regulatory activities is usually associated with transcriptional repression.[371,372,374]

Human CHD3 was first discovered, together with CHD4, as an autoantigen recognized by anti-Mi-2 antibodies in 15-20% of dermatomyositis patients.[375–377] Dermatomyositis is an inflammatory disease affecting mainly skin and muscle tissue, and is associated with increased risk (~25%) of developing cancer.[378] Genetic somatic alterations in *CHD3* have also be identified in cancers from patients without dermatomyositis.[379,380] Together, this evidence indicates a possible involvement of this gene in human disease. Furthermore, autosomal dominant constitutional mutations in other CHD genes have already been shown to cause disease: *CHD2* has been linked to intellectual disability with mild dysmorphism [381,382] and to childhood-onset epileptic encephalopathy which presents with epilepsy and cognitive delays,[383] *CHD7* causes CHARGE syndrome (OMIM #214800) which is characterized by growth retardation and

multiple congenital defects,[384] and *CHD8* is thought to provide susceptibility to autism in combination with increase head size and tall stature.[385]

Interestingly, the DECIPHER database lists five patients from the Deciphering Developmental Disorders study with rare *de novo* variants in *CHD3*: four missense and one nonsense. One of the missense variants reported, p.(Arg985Trp), affects the same amino acid that is modified in our patient (NP_001005273.1). None of these variants (or ours) have been previously reported in dbSNP, ExAc, ClinVar, COSMIC, or the Exome Variant Server. Unfortunately, the phenotypic information associated with these patients on DECIPHER is unspecific (listed as "abnormalities" of various organs/tissues). Obtaining further phenotypic information on these five patients to determine whether they show any phenotypic overlap with our patient will be extremely valuable in investigating the pathogenicity of rare variants in *CHD3*. Alternatively (or additionally, as another line of evidence), functional investigations may be required.

Importantly, *CHD3*'s most closely related gene *CHD4*, was shortlisted as a strong candidate for causing a Mendelian phenotype in humans based on information derived from population genetics and mouse knock-outs.[386] Mechanistically, Chd4 has been shown to interact directly with Ezh2 in mouse neural progenitor cells, and to be required for Ezh2-mediated silencing of the astrogenic marker gene *GFAP*, together preventing premature onset of gliogenesis; this study supports a role for Chd4 as a context-dependent co-factor of PRC2.[387] Very recently, in a study published online in September 2016, Weiss *et al*.[388] described five unrelated individuals with *de novo* missense mutations in *CHD4* that presented with developmental delay, intellectual disability, macrocephaly and distinct facial dysmorphisms, among other features. Their study included cell-based functional assays that supported pathogenicity for the variants described in *CHD4*. The photographs published in this paper demonstrate features that significantly overlap with the clinical features observed in our case 92. As such, it is highly plausible that this variant in *CHD3* could be causative of her overgrowth phenotype. This variant will remain our strongest candidate until further cases are identified through collaborations.

## 4.5 Conclusions from singleton exome sequencing strategy

### 4.5.1 Challenges in variant interpretation for novel disease genes

The main challenge and concern when using exome sequencing to determine the cause of disease is accurate variant interpretation;[337,340,389,390] this is partly why we were only able to conclusively diagnose one patient using this high-throughput method. In the field of rare diseases, and particularly when searching for new causes of sporadic developmental syndromes, we are actually fortunate that we can classify individuals as "affected" or "unaffected" with a high degree of certainty, because many of these disorders manifest through congenital anomalies and/or complications in early childhood.[292] Further, the assumption is that such rare disorders are caused by rare variants of strong effect on protein function, which is why *de novo* mutations have a higher prior probability of causing disease in families where only one individual is affected and no consanguinity is reported (although caution is warranted because every individual carries several non-synonymous *de novo* variants and they do not all cause disease).[391] Thus, variants observed frequently in the greater population can be excluded from analysis, and variants unique to each proband can be prioritized (reducing the list from ~20,000 to ~400 variants).[1,343] Yet, despite these advantages over studying complex and/or adult-onset disorders, we are still left with a high number of variants to interpret.[392]

Guidelines for variant interpretation in the clinical setting have evolved over the past few years, and must be followed strictly, with a specific set of rules that make the interpretation of variants conclusive at any given time.[341,393] For example, novel variants in known disease genes in patients with overlapping phenotype are still classified as "probably pathogenic", even when they are *de novo* and predicted to be damaging. Similarly, variants in genes that haven't been previously associated with disease, even if they appear to be strong candidates (based on known protein function, population genetics and pathogenicity prediction scores), cannot be investigated further and must be classified as VOUS (some laboratories may report them as "GUS" for genes of uncertain significance).[393] To address these limitations, it is recommended that exome/genome data from unsolved cases be re-analyzed routinely (generally every six months to a year), to allow for incorporation of novel gene discoveries and new classifications of variants as reported in the scientific literature and/or public databases.[3] Patients should be informed in advance whether this re-analysis will occur.[2]

In contrast, guidelines within the research setting are less rigid to allow for the discovery of new gene-disease associations. Researchers have more tools (as well as time and money) at their disposal to investigate the potential pathogenicity of variants, both for novel variants in known disease genes, and for variants in genes not previously associated with disease. However, this also means that investigations done on a research-basis could in theory be never-ending. In reality, these are usually limited by the budget available, so it falls on the researcher to prioritize which investigations to pursue. Importantly, these investigations should require a high standard of evidence, and interpretation of novel variants without strict guidelines must not compromise accuracy. Supporting evidence can come from population genetics, bioinformatic predictions, and/or functional experiments using model organisms. Some of the strategies and resources available are discussed in the following sections.

#### 4.5.1.1 Using the knowledge of population genetics available in public databases

Population genetics is one of the most powerful tools for variant interpretation. In our quest for accuracy, several databases must be searched in order to interpret a single variant,[1,342,390] as described in Chapter 2, section 2.3.2.1.5. This is particularly important for missense variants, which may or may not have an effect on protein function; this effect is hard to predict even when combining several bioinformatic tools.[337,394–396]

Variants that are not reported in dbSNP, ExAc or the Exome Variant Server can be considered extremely rare in the general population, and have in theory a higher probability of being pathogenic for developmental disorders. However, it is important to remember that not all ethnic groups are well represented within these databases,[269] which is an important limitation of these studies, and therefore the population frequencies reported should not be used uniformly for all patients; the possibility of a variant being enriched within a specific population (with a higher frequency only within that population that is masked when looking at mixed populations) should be taken into account.[268] Variants described in the COSMIC database have a previous link to human cancer (in somatic state), which may or may not reflect pathogenicity of the variant in its constitutional state because cancers have a much heavier mutational load and their development is usually not explained by a single genetic alteration. In other words, the somatic variant may be a passenger mutation rather than a driver mutation, and thus have no involvement in disease pathogenesis. Variants reported in ClinVar are constitutional (with the exception of some mosaic

alterations), and will usually have more detailed information regarding their classification; but this information needs to be interpreted carefully, using their internal classification system [397] as well as our own critical thinking, due to the multitude of data submitters involved. Variants reported in DECIPHER and LOVD-specific databases are also usually constitutional, and again have in theory a higher probability of causing disease; however, it is important to assess the robustness of the evidence provided, as well as to determine whether the phenotype that is associated with the reported variant shows sufficient overlap with that of the patient being tested.

These public databases were originally constructed to compile information on human genetic variation in the wider population and make it publically accessible, which should support more straightforward variant interpretation, particularly for rare developmental disorders that are ascertained in early years. However, without strict curation methods, these databases gradually accumulate inaccurate and/or contradictory information that ultimately lengthens the process of variant interpretation.[285,390] For example, dbSNP was originally intended to reflect common genetic variation across healthy populations of various ancestries,[347] including all the variants identified through the 1000 Genomes Project.[268] However, nowadays, rare variants are also reported in dbSNP,[343] and linked to ClinVar when clear evidence of pathogenicity is available. This means that the simple indication that a variant has an associated "rs number" (meaning that it is reported in dbSNP) no longer constitutes evidence to exclude it from the candidate list of causative variants for rare disorders, as was previously done in a systematic manner using bioinformatic filters. Another example is the ExAc data, which has recently been incorporated into dbSNP. On a closer look, many of the variants reported by ExAc have only been observed in a single individual (~54% of all variants from ExAc, according to Lek *et al*.[269]) and have not been validated by Sanger sequencing. Although these variants were considered to be "high-quality variants" through bioinformatic stringency,[269] in reality, these can be either true extremely rare variants or false positive hits. We must note that the majority of individuals whose variants were reported by ExAc are not healthy individuals; many have adult-onset disorders, which means that these variants may be associated with disease. However, they are not (in theory) linked to pediatric developmental disorders, which is why these data can still be used for variant interpretation in rare disease patients. Nonetheless, these variants do not reflect common genetic variation, adding to the pool of "non-SNPs" deposited into the dbSNP database.

These examples illustrate both the usefulness and limitations of using public databases for the interpretation of variants. Despite many efforts, a definitive database that is highly curated will be difficult to achieve,[390] which means that we must keep querying a variety of databases to interpret each variant; ideally, new bioinformatic tools should be able to combine all the information in a single file. Furthermore, we need to make sure that previously reported rare variants do not mask the "uniqueness" of variants upon bioinformatic filtering (for example by filtering out a variant reported in dbSNP because it was observed in another affected individual),[343] which is why it is still important to look carefully at all variants within prior candidate genes. In conclusion, evidence from population genetics should be meticulously assessed before using it to include or exclude any variant from the candidate list, and this evidence should be interpreted in conjunction with information derived from phenotypic and/or functional analysis.

#### 4.5.1.2 The power of informative phenotyping

As mentioned earlier, the field of rare diseases relies heavily on detailed phenotyping for the interpretation of WES results. However, this strategy involves reviewing clinical charts and a myriad of other documents,[292,390] as we have done, which is extremely time-consuming, and then manually curating the variant list in relation to the phenotypic information, which is again extremely time-consuming. In an attempt to address this, bioinformatic tools have been developed to prioritize variants from each patient in relation to their phenotype in an automated fashion.[389,398,399] Many of these make use of Human Phenotype Ontology (HPO) terms, which were created in an effort to standardize phenotyping across the field (and even across species).[400] However, these tools can only link HPO terms to genes that have associated HPO terms themselves,[389,392] so many genes are left out of the automated prioritization and/or wrongfully excluded based on lack of information, thus limiting the opportunity for novel gene-disease associations. Furthermore, the phenotypic information from a single patient is often insufficient to determine whether a particular variant could lead to the observed phenotype, or to prioritize between several promising candidate variants.

Another well-recognized strategy to determine pathogenicity of variants in a gene that hasn't been previously associated with disease is to find two or more unrelated patients with the same disorder who have rare variants within this gene.[1,3] Because one investigator may not come

across several patients with the same rare disease within a single genetics centre, web-based data-sharing platforms have been created to connect researchers from different centres who may have patients with similar phenotypes associated with variants in the same gene or pathway. These tools also make use of HPO terms for standardized annotations. Such platforms include Canada's PhenomeCentral,[401] which is now integrated into Matchmaker Exchange together with other international databases and programs.[402] These collaborative platforms are expected to speed up new gene-disease associations and, as a result, aid the interpretation of variants identified in novel disease genes flagged by independent researchers (i.e. variants which would require other lines of evidence to predict pathogenicity when observed in isolated cases). Although we only identified two unrelated patients with rare variants in *EED* in this study, the phenotypic overlap was compelling and heavily supported by population genetics and the known function of EED.

#### 4.5.1.3 Functional assessment to predict pathogenicity of variants

When population genetics and phenotypic information are not sufficient to determine the causative disease variant/gene, we must investigate other lines of evidence. Another powerful tool for variant interpretation is the knowledge of what the protein encoded by that gene does at the biological or functional level,[337,343] and how this function may relate to the observed phenotype. *In vitro* and *ex vivo* assays (for example using patient-derived cell lines) can be very useful for such investigations, but they do require some prior knowledge of protein function and/or structure in order to design the assays and determine the best output of such assays. Therefore, functional experiments are extremely limited by the fact that a concrete protein function has not yet been determined for a large number of genes,[1] and further investigations are required to determine the function of lesser known proteins/genes.

##### 4.5.1.3.1 Bioinformatic predictions

Given that there are numerous variants to interpret within each exome, the ideal functional assessment should be fast, cost-effective, and scalable to vast amounts of data. In an attempt to address this need, recent bioinformatic efforts have focused on developing *in silico* prediction models to infer on plausible functional roles for uncharacterized genes. Some models explore relationships between proteins, with the assumption that functionally related proteins (for

example within the same pathway or acting within the same complex) may lead to a similar disease phenotype when disrupted.[403] However, human protein and gene pathways are intricate and extremely interconnected, which makes it hard to trust these predictions for variant prioritization and interpretation. Other models extract what is known about a particular gene in other organisms to make functional predictions, with the assumption that genes causing disease are essential and thus likely to be well conserved across species.[404] Overall, these tools may help prioritize candidate genes, but they lack the robustness required to definitively link a novel gene to a phenotype; further assessment should be carried out using *in vivo* models.

#### 4.5.1.3.2 Model organisms

In order to assess the function of human genes within a live complex organism, we can generate animal models where orthologous genes are mutated or knocked-out. Among these model organisms, the mouse is the mammalian model most commonly used to study developmental disorders. Because much work and expertise are required to generate animal models, there has been a strong incentive for collaborations between clinical researchers and model organism specialists.[1] With this in mind, the International Knock-out Mouse Consortium was formed, and these collaborative groups are attempting to characterize the function of every mouse gene through the construction of mutant ES cell lines and subsequent generation of mouse lines (if viable).[405] This initiative has confirmed that many genes causing Mendelian disorders through heterozygous mutations in humans will lead to embryonic or perinatal lethality in knock-out mice,[405] similarly to what had been observed for *EZH2*, *NSD1* and *EED*. Interestingly, the most common phenotypes observed in mouse embryos lacking such essential genes (assessed at various time-points during development) were growth and developmental delay, followed by cardiovascular abnormalities, craniofacial malformations and defects in limb development, suggesting that loss-of-function of these genes in humans could cause similar congenital defects.[386] Importantly, although they may help us determine the prior probability of causality for novel disease genes, these findings must be interpreted with caution because different mechanisms are involved (heterozygous vs. homozygous mutations). Yet, some of these phenotypes were also observed in heterozygous mice, proposing that they may not be exclusively due to loss-of-function, and therefore that different types of alterations within these genes could lead to disease. Curiously, many of these mouse essential genes have not yet been associated

with disease in humans, so they represent good candidates for rare developmental disorders that still remain undiagnosed.[386] Again, this evidence should be looked at with reservation: even if the protein is highly conserved across species, its function may be different and therefore mutations in the encoding genes may lead to different phenotypes.[291,392]

#### 4.5.1.4 Final remarks regarding variant interpretation for novel disease genes

Within a novel disease gene, the "perfect" variant to call pathogenic for an isolated presentation of syndromic developmental defects would be: *de novo*, not previously reported in any database, and predicted to disrupt protein function, preferably a function that could clearly lead to the observed phenotype. Ideally, this would be supplemented by evidence that the patient's phenotype is consistent with the phenotype of other individuals with mutations in the same gene. However, this does not happen very often.

There is simply no shortcut for fast and accurate variant interpretation, and this step still requires extensive manual curation. Naturally, this manual curation is subject to personal judgment and can lead to differing interpretations across laboratories, which could subsequently have a significant impact on patient case.[393,397] This is why it is so important that scientists deposit their findings into public databases.[3,390] Current practice focuses on reporting causative variants only, but reporting variants conclusively disproven to be associated with a particular phenotype is equally important. Data sharing is essential for the progress of rare disease research, and will save time in future interpretations as well as promote much-needed consistency of variant interpretation across centers.[390]

Additionally, the need for manual curation is probably the main reason why the cost of WES as a service cannot decrease much further, and it will take time before we can move towards a more automated system. Information derived from population genetics can inform on the rarity of each variant as well as on the mutational burden of each gene to estimate pathogenicity, a concept well explored by ExAc. The incorporation of phenotypic information for variant prioritization is also well underway, with novel tools such as Phen-Gen [398] and VarElect [406] being developed by numerous independent groups. Furthermore, access to animal data such as that generated by the International Knock-out Mouse Consortium should reduce the necessity for generating new (and costly) models to investigate each gene. Recently, Dickinson *et al*. surveyed this data to generate a list of genes not yet associated with human disease that are essential in

mice and also predicted to be highly intolerant to mutations in humans (using ExAc scores); these genes represent strong candidates to explain undiagnosed developmental syndromes.[386] This example proves that combining multiple high-throughput datasets can support variant interpretation, but the real challenge lies in developing a bioinformatic pipeline capable of incorporating all these different datasets into a single platform. Although some data analysis software will allow for manual incorporation of multiple datasets, this is still not possible in a fully automated manner. Until then, the field will continue to rely on combined efforts from multidisciplinary teams of highly trained professionals to carry out variant interpretation.[1,2,390]

Notably, incidental findings, a major concern in WES studies, are not discussed here. This is because ours is a research study, and we only looked at rare variants unique to each patient in an effort to prioritize potential causes of overgrowth.[3] We did not specifically look at other (non-unique) variants, including those within the 56 genes listed in the guidelines from the American College of Medical Genetics and Genomics (ACMG);[407,408] this strategy was explained at the time of consent and is in alignment with the position statement from the Canadian College of Medical Geneticists (CCMG).[2]

### 4.5.2 Other limitations of our gene discovery strategy

#### 4.5.2.1 Selection of patients to sequence

We chose to sequence singletons who were reported to have a "Weaver-like" phenotype and overall severe clinical presentation at the time of enrollment in our study. Our goal was to prioritize cases with distinct phenotypes for higher probability of novel scientific discoveries, while also considering clinical urgency regarding the potential impact of our findings on the patient's care.[3] Additional phenotypic data were collected later, while WES was ongoing, knowing that it would be necessary for the variant interpretation phase. This supplemental data revealed that, despite being referred for Weaver syndrome, some patients actually demonstrated few "Weaver-like" features and had a rather unspecific overgrowth phenotype. For one patient (case #68), the overgrowth observed prenatally and immediately after birth had resolved, and the phenotype had since become evocative of an undergrowth syndrome instead (with growth parameters significantly lower than that of his twin). Therefore, the ten patients we selected may not have been the best patients to sequence in order to identify novel overgrowth genes. For

future studies, we recommend collecting as much phenotypic information as possible at the time of enrollment into the study, or at least before selecting patients for WES.

Ultimately, despite our intention of selecting patients with significant phenotypic overlap, it is not surprising that we did not find a strong candidate gene mutated in two (or more) unrelated individuals through WES. Nonetheless, it is remarkable that we were able to establish one novel gene-disease association within our WES data. One similar overgrowth gene discovery was reported by Tatton-Brown *et al*., regarding *DNMT3A*,[160] but this group sequenced ten trios rather than just ten singletons. Similarly, *SETD2* was identified as a novel overgrowth gene in two patients through targeted sequencing of 16 singletons,[161] suggesting that selecting another sub-group of patients for us to sequence (for example ten more singletons) could be sufficient to identify novel variants within the same gene in at least two unrelated patients. It is also likely that we would identify more patients with novel mutations in genes recently shown to cause overgrowth (like *DNMT3A* and *SETD2*), rather than identifying another novel overgrowth gene, considering that competing groups have much larger cohorts than we do (~150-200 patients).

#### 4.5.2.2 Variants missed in our analysis strategy

As expected, interpretation is more straightforward when a similar phenotype has already been associated with other variants in the same gene.[1,341] With this in mind, we excluded variants in genes associated with phenotypes distinct from those observed in our patients. Yet, we know from recent discoveries that differing phenotypes can be associated with different mutations in the same gene,[1,291,337] and thus we may have excluded good candidates from our analysis.

Moreover, some genes have been associated with reciprocal growth phenotypes. For example, homozygous mutations in the natriuretic peptide receptor-B (*NPR2*) were known to cause acromesomelic dysplasia Maroteaux type (AMDM, OMIM #602875), a skeletal dysplasia leading to extreme short stature,[409,410] and carriers of the same loss-of-function *NPR2* mutations (in heterozygous state) were shown to have non-syndromic short stature.[410–415] Later, case reports of missense mutations in *NPR2* leading to tall stature emerged; these were predicted to be gain-of-function mutations through functional studies.[410,416–418] Another example of relevance here is that duplications involving *NSD1* have been shown to cause an undergrowth syndrome characterized by short stature, microcephaly, delayed osseous development, intellectual disability, and occasional mild dysmorphism;[419–423] this represents a reciprocal presentation of

that observed in patients with Sotos syndrome, which is caused by haploinsufficiency of *NSD1*. In addition, 24 cases with copy number variants involving *EZH2* are described in the DECIPHER database, and general phenotypic information was available for nine of them: four deletions and four duplications are associated with short stature/undergrowth, while one duplication is associated with tall stature/overgrowth. Together, these examples suggest that bidirectional growth dysregulation can be caused by dosage imbalances of the same gene, and that reciprocal imbalances in protein activity attributable to other mutational mechanisms (such as constitutive activation or inactivation of enzymatic activity) might also cause reciprocal growth phenotypes. Therefore, genes already reported to cause undergrowth (which we excluded in our variant prioritization) should probably be considered as plausible candidates in future investigations of overgrowth syndromes.

Another important thing to remember is that we have carried our analysis with the assumption that the phenotype would be explained by a single, rare, *de novo* variant. However, recent reports have shown that some syndromic presentations that do not match the typical phenotype of that syndrome can sometimes be explained by two separate genetic causes, and therefore the unusual phenotype observed is actually a combination of two separate phenotypes (called "blended" phenotype).[1,3,338,339,424–426] This possibility was not addressed in our WES analysis and should be considered upon re-analysis of the WES data.

Finally, WES does not detect non-coding variation, as well as copy number variants, structural variants, trinucleotide expansion repeats, mitochondrial DNA variants, and changes at the epigenetic or RNA levels.[291,427] WES is also unlikely to detect somatic variation, unless carried out at high coverage. Since a considerable amount of genetic (and epigenetic) variation is missed using WES, other experimental and analysis strategies will be required to determine the cause of disease in some undiagnosed patients.[1,291]

### 4.5.3 Alternative strategies to consider

#### 4.5.3.1 Trio-based exome sequencing

The most obvious alternative strategy to consider here is trio-based WES, where the patient's DNA and the DNA from both parents are each sequenced in order to filter out inherited variants all at once (when looking for *de novo* variants as in our study; alternative filtering may also be applied to compile a list of genes with variants in the compound heterozygous or

homozygous state, for the investigation of recessive disorders). This strategy is incredibly powerful in shortening the list of candidate variants from hundreds to only a handful,[343] and is currently the most used strategy for clinical WES.[1,341] Trio-based WES is time- and cost-effective not only because it shortens the list of candidate variants, but also because it reduces the number of independent validations required (indeed, many of the candidate variants we selected for validation turned out to be inherited).

With trio-based WES, the chances of finding a good candidate are higher, but still not guaranteed.[3] This strategy also allows for filtering of variants without relying as heavily on public databases,[292] which are known to have errors (as discussed earlier) and to underrepresent variation of certain ethnicities.[342] However, we must remember that this approach is much more costly on the technical side (three times the cost), even though the interpretation cost based on time required for analyzing the results and classifying variants is comparable. This strategy is more appropriate for investigating isolated cases or very small cohorts, when no similar cases are available for comparison and the strategy of "phenotypic overlap" cannot be applied.

#### 4.5.3.2 Whole genome sequencing

We could also consider using "whole"-genome sequencing (WGS). The exome represents only 1-2% of our genomes,[1] so WGS could uncover non-coding causes of disease across the remaining 98% of the genome, as well as some copy number variants (CNVs) and structural aberrations, depending on the bioinformatic pipeline.[291] Currently, the incremental benefit of this strategy in rare diseases (as compared to exome sequencing) is unclear, because most rare diseases diagnosed so far are caused by mutations in protein-coding regions [291] and large CNVs are usually ruled out by chromosomal microarray. Yet, about half of the ~7,000 Mendelian disorders remain undiagnosed despite significant efforts using WES.[291]

As mentioned earlier, there are many genes predicted to have an effect in humans that have not yet been associated with disease, suggesting there are still many gene-disease associations to be established.[291] Interestingly, WGS has been reported to have a better coverage of the exome than WES, so this strategy may also uncover coding variants missed by WES.[3,337] This is important to consider since most laboratories use the same commercially-available exome capture kits, and thus the same parts of the exome are consistently under-sequenced. WGS may help reduce these gaps in coverage and establish new gene-disease associations that had been

missed due to technical limitations. WGS may also detect indirect disruptions of protein-coding genes, for example through alterations in regulatory regions, that would have been missed by looking exclusively within the exome.[3,291]

However, with more data come more challenges in bioinformatic analysis and variant interpretation [337] The role of most non-coding variants and the genes to which they relate to are still widely unknown,[1] and the possibility of incidental findings is much greater. The computational and man power required to analyze these data are also exponentially greater, along with higher costs.[3] WGS may eventually become the standard method for rare disease research, once gene discoveries using WES become too scarce and the costs become comparable to WES,[3] but for the time being WES is still the most cost-effective method.

# Chapter 5: Discussion

In my thesis work I have identified and described pathogenic mutations in *EZH2* and *NSD1* causing the known Weaver and Sotos syndromes, and expanded the phenotypic spectrum of Weaver syndrome (WS) (Chapter 2). I have also determined that the *EZH2* mutations observed in WS patients are likely to impair PRC2-mediated H3K27 methylation, but that this impairment does not correlate with the severity of the phenotype observed (Chapter 3). Finally, I have discovered that constitutional mutations in *EED* can also cause overgrowth (Chapter 4); this gene encodes another PRC2 member and chromatin regulator, as hypothesized at the start of this project.

## 5.1 Overall diagnostic rates

The full progression of the diagnostic strategy, which included both targeted sequencing of known disease genes and the use of exome sequencing (WES) for novel gene discovery, is presented in Figure 5-1.

When including variants classified as "probably pathogenic" and the two cases that received a different diagnosis externally, the final diagnostic rate within our cohort by June 1st 2016 was 33% (22/66). Although similar frequencies have been reported for diagnosing well-characterized cohorts via WES (e.g. 32.5% for neurodevelopmental disorders [428]), we only used targeted gene testing for most of our patients, which means that our own diagnostic rate is actually quite high. This total includes a 17% rate using direct sequencing of *EZH2* and 11% using direct sequencing of *NSD1* (including validations of previously known positive results), for a total diagnostic rate of 27% (18/66) using Sanger sequencing. This diagnostic rate reflects a good pre-selection of cases with Weaver and Sotos syndromes based on detailed phenotyping, and is consistent with previous reports (see section 2.4.3).

In contrast, only one individual was diagnosed via WES among the 46 patients left to diagnose, for a success rate of 2%. This is due in part to the selection of samples for WES without sufficient phenotypic overlap to define a novel syndrome (as discussed in Chapter 4, section 4.5.2.1), but also to the heterogeneous clinical presentation observed within the remaining undiagnosed cohort (as discussed in Chapter 2, section 2.4.1.4). This phenotypic

heterogeneity strongly suggested genetic heterogeneity in the underlying causative genes, making gene discovery more challenging. Nonetheless, we were able to identify *EED* as a novel overgrowth gene in one patient, using the WES data of only eleven singletons, for a discovery rate of 1:11 that is also comparable to other studies (as discussed in Chapter 4, section 4.5.2.1). What is remarkable is that we subsequently identified a different *EED* mutation in a second unrelated patient, among only 35 patients left to test, particularly considering the small coding region of this gene (and thus the reduced "mutational target space"). Together, these findings reinforce the power of detailed phenotyping in achieving a molecular diagnosis for patients with rare overgrowth syndromes.

Of the 44 patients that remain undiagnosed, 13 (20%) have a VOUS in a known overgrowth gene that requires further investigation. This leaves us with 31 patients overall (47%) that are lacking a candidate variant to explain their phenotype. Within this undiagnosed cohort, we are likely to have individuals with undetected alterations in known disease genes, individuals with unusual presentations of known overgrowth syndromes (including cases that may be explained by two separate genetic causes, as mentioned earlier), and individuals with extremely rare new syndromes. There is still potential for new gene discoveries, but the likelihood of again finding two (or more) unrelated patients with mutations in the same novel gene is much lower. This is supported by the fact that competing groups with larger cohorts were also fast at identifying a single novel cause of overgrowth, but have not been quick to publish subsequent gene discoveries. Additional collaborations will be needed,[343] and/or we will need to carry out functional studies to establish the pathogenicity of new candidate genes identified in "N of 1" cases.

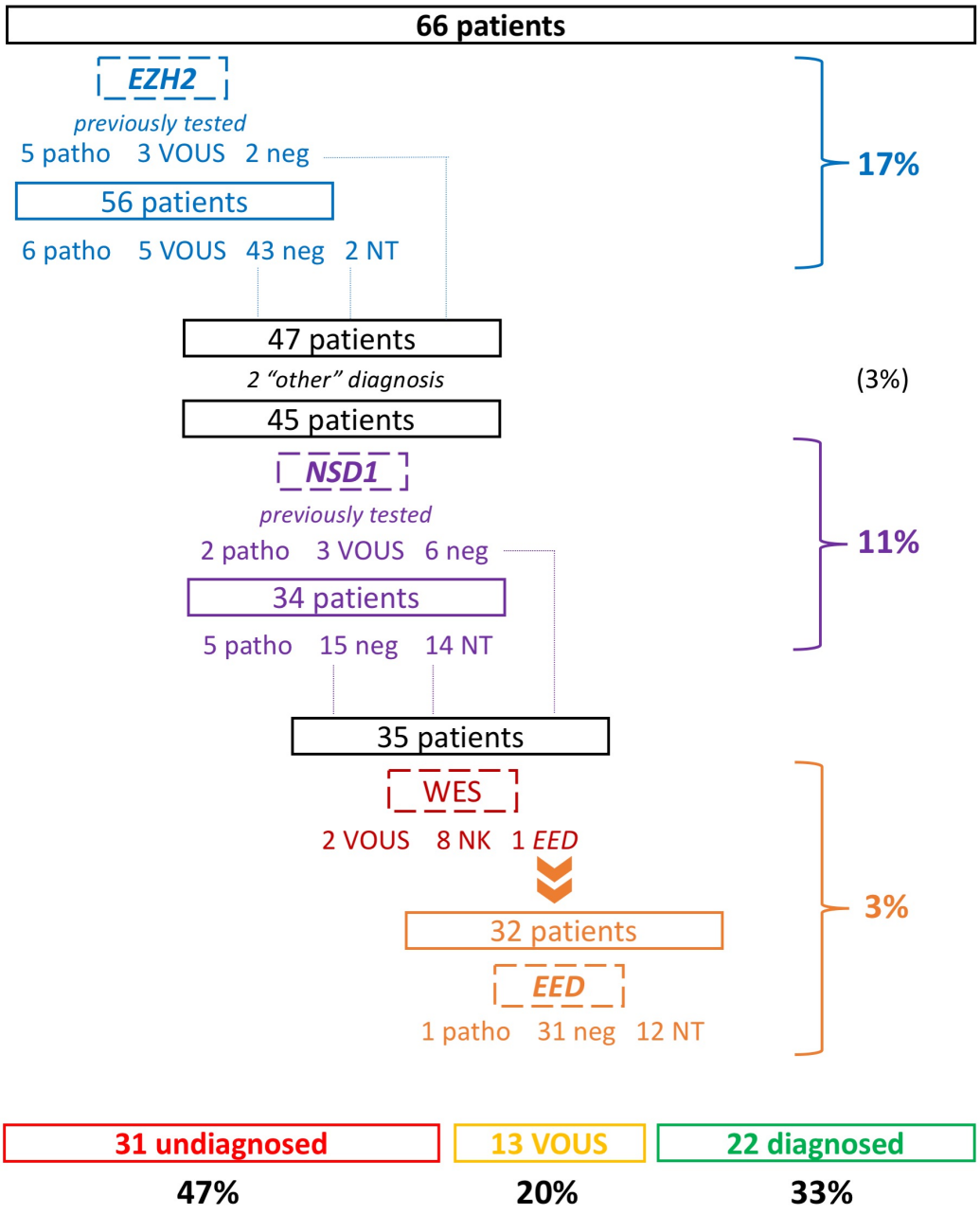



**Figure 5-1: Overall diagnostic rates from our study.**

patho = pathogenic or likely pathogenic variants; VOUS = variants of unknown significance; neg = negative; NT = not tested; NK = "not known" for unresolved exomes. Percentages on the right represent patients diagnosed at each stage of investigation, calculated out of the total number of patients (N = 66).

## 5.2 EZH2 vs. EED

Although expected, it is interesting that mutations in both *EZH2* and *EED* can cause overgrowth syndromes. The two proteins are essential members of PRC2; so naturally, disrupting either protein could impair PRC2-mediated H3K27 methylation, which would theoretically lead to the same disease phenotype. Yet, we do observe some differences between *EZH2* and *EED* mutation positive overgrowth patients, as discussed in Chapter 4 and summarized in Table 5-1 below.

| | **Weaver syndrome (WS)** | ***EED*-related overgrowth** |
|---|---|---|
| **General** | Overgrowth<br>Tall stature<br>Accelerated osseous maturation<br>Intellectual disability (milder) | Overgrowth<br>Tall stature<br>Accelerated osseous maturation (less advanced)<br>Intellectual disability |
| **Head** | Macrocephaly (~50%)<br>Large bifrontal diameter<br>Flat occiput | Macrocephaly<br>Large bifrontal diameter |
| **Face** | Round (early years) | Round (early years) |
| **Eyes** | Hypertelorism | Eye abnormalities (including myopia)<br>Full thick eyebrows |
| **Ears** | Large and fleshy | Large |
| **Chin** | Prominent chin crease ("stuck-on" chin)<br>Receding jaw (micro/retrognathia) | Prominent chin crease ("stuck-on" chin)<br>Receding jaw (micro/retrognathia) |
| **Others** | Cerebral malformations<br>Cancer predisposition (5-10%)<br>Umbilical hernias<br>Excessive "doughy" skin<br>Hoarse low-pitch cry<br>Long philtrum<br>Wide nasal root<br>Congenital cardiac anomalies (less frequent) | Large hands<br>Long slender fingers (arachnodactyly)<br>Umbilical hernias<br>Increased pigmented nevi<br>Skeletal abnormalities |

**Table 5-1: Clinical presentation of Weaver syndrome vs. *EED*-related overgrowth.**

Remarkably, these phenotypic differences (noting that we only have two cases of *EED*-related overgrowth on whom to base these remarks) are only apparent in adulthood, as discussed in Chapter 4, section 4.3.4.3. Indeed, both of our *EED* positive patients had typical WS features in early childhood, which make *EED*-related overgrowth likely undistinguishable from WS at an early age. The exception may be with regards to skeletal abnormalities, which were significant in both our *EED* positive patients and also in *Eed* knockout mice;[222] thus, we would recommend a

detailed assessment of spinal column structure and stability using X-rays, followed by MRI if neurological deficits are present. As more patients are diagnosed, we will be able to determine whether a clinical distinction between *EZH2* and *EED* mutation carriers is important, particularly considering that a distinction between WS and Sotos syndrome has already been proven useful for clinical management. One possible difference may lie in tumour surveillance: as mentioned earlier, our two *EED* positive patients, who are both in their early 30s, have not yet developed malignancies. So it is possible that patients with constitutional mutations in *EED* are less likely to develop cancer in childhood than patients with constitutional mutations in *EZH2*, making a concrete molecular diagnosis important at an early age.

The observed differences in phenotypic appearance suggest that disruption of PRC2-mediated H3K27 methylation is unlikely to be the only mechanism contributing to disease, which would also support the conclusions drawn from our functional studies presented in Chapter 3. As such, we need to consider other roles of EZH2 and EED. We know that both proteins are required early in mouse embryonic development, but *Ezh2* expression begins earlier (and lasts longer), suggesting that Ezh2 plays additional roles in development that do not require Eed.[221] Further, EZH2 is the catalytic enzyme that mediates H3K27 methylation, whereas EED does not have catalytic function itself and instead is thought to recognize H3K27me3 to recruit both PRC2 and PRC1 to propagate repressive marks;[203,429] so modifying each protein could affect gene silencing differently, possibly by altering the distribution of silencing marks across the genome, which would in turn have different effects on downstream expression of target genes. Finally, it is thought that EZH2 activity can be partially compensated by its paralog EZH1.[430,431] Any of these characteristics, individually or together, could contribute to the physical differences observed; some functional investigations are proposed later, in section 5.4.2.

Furthermore, according to the ExAc database,[269] *EED* is expected to be more tolerant of missense mutations than *EZH2* (z-score of 2.69 compared to 5.45, both being highly intolerant to loss-of-function), and yet constitutional mutations in *EZH2* have been observed more often (and were discovered earlier) than mutations in *EED*. We must remember that *EED* has a smaller coding sequence (2,476 bp compared to 2,723 bp, but encoding only 441 amino acids compared to 751 amino acids), and thus the "mutational target space" is smaller. It will be interesting to see how many cases of *EED*-related overgrowth are reported in the next few years, and what the full phenotypic and mutational spectrums of this novel syndrome will look like.

Lastly, based on this discovery, we would highly expect mutations in the third essential member of PRC2, *SUZ12*, to cause another "Weaver-like" overgrowth syndrome. *SUZ12* is actually larger than *EED*, with 4,517 bp (NM_015355.3) encoding 739 amino acids (NP_056170), predicted to be similarly intolerant to missense mutations (z-score of 3.68), and highly intolerant to loss-of-function. Further, like the other two genes, *Suz12* is essential for mouse embryonic development, with knockout mice being embryonic lethal (though heterozygous mice do survive).[223] Somatic mutations in *SUZ12* have been described in various cancers,[233] but constitutional SNVs or indels in *SUZ12* have yet to been associated with human disease. Importantly, some neurofibromatosis patients with microdeletions surrounding the *NF1* gene present with overgrowth and Weaver-like dysmorphism, and it has been hypothesized that these features are caused by haploinsufficiency of *SUZ12*, which is located within the deleted region.[432–434] This suggests that mutations in *SUZ12* may in fact cause a Weaver-like overgrowth syndrome that has not yet been ascertained.

## 5.3 Conclusions

### 5.3.1 Strengths

As demonstrated by our high diagnostic rate, the use of detailed phenotyping represents the main strength of this thesis work. This strategy is highly recommended for other investigations seeking to establish new gene-disease associations for rare diseases and other genetic disorders.

### 5.3.2 Limitations

#### 5.3.2.1 Availability of samples and patient data

One of the main limitations of this project, as well as other projects investigating rare diseases, is the limited sample size. Due to the rarity of WS, we had to recruit patients from all over the world. This geographical distance translated into DNA samples of variable quality and also made it virtually impossible to collect fresh tissue on a significant subset of patients, which means that other lines of evidence that would be helpful for understanding the pathogenesis of WS (such as measuring EZH2 or H3K27me3 protein levels directly on patient samples) were not an option.

Similarly, the level of detail and accuracy of the phenotypic information provided was dependent on the level of expertise and engagement of the referring physician. In an effort to "even out" the information received for each patient and assure that essential phenotypic data were collected, we provided a detailed list of documents and consultation work-up recommendations, if requested. However, we may have missed important features that have not yet been described in overgrowth syndromes. In addition, we found that providing our own phenotypic table as a point of reference for the information we were seeking, or even asking healthcare providers to complete this table, was not helpful in most cases because physicians often do not have time to carry out such detailed descriptions. Furthermore, we also accepted samples and clinical information directly from families; they were eager to help but unaware of which information would be helpful for our study, so we often received vast amounts of clinical records that were extremely time-consuming to review and that did not provide the phenotypic details that we needed. Ultimately, we have limited control over the quality of phenotypic information available for each patient, and therefore we will always have gaps in our database.

#### 5.3.2.2 Technical limitations

##### 5.3.2.2.1 Sequencing studies

As discussed in Chapters 2 and 4, genetic variation beyond coding SNVs and small indels would not have been detected in this study, and a myriad of other techniques would be required to provide a complete genetic evaluation of our patients. Interestingly, we were able to detect a somatic *EZH2* p.(Glu441=) variant in proband 10 at a 15-20% level of mosaicism (Figure 2-3). However, it is likely that other cases of mosaicism may have gone undetected, particularly when low-quality DNA was used that produced noisy Sanger peaks. Indeed, although the sequencing protocol was optimized, there was often some residual background observed on the Sanger sequencing traces, which did not interfere with the detection of fully heterozygous variants, but could easily interfere with detection of mosaicism. This protocol would definitely have to be further optimized to progress to clinical-grade testing. In addition, we should consider testing additional tissues from each patient, although this would be difficult considering the geographical distance (as discussed above). These limitations are important to consider given that, in the last few years, somatic mutations have been shown to play a much bigger role in disease than previously appreciated.[6,435] Of particular interest, two mosaic *NSD1* intragenic

deletions have been reported: one patient was classified as having typical Sotos syndrome,[436] but the other did not show the characteristic facial appearance and did not have overgrowth,[437] adding an additional layer of complexity to the clinical heterogeneity observed with alterations of a single overgrowth gene. Further, mosaic (epi)genetic alterations were detected in tongue tissue from three Beckwith-Wiedemann syndrome patients who had received negative results in routine screening using blood samples.[438]

Individuals found to be mosaic for mutations in *EZH2* may or may not show WS features, and may or may not develop malignancies. It is likely that some parents ascertained as being unaffected may actually carry mosaic mutations, which may also be present in the gonads and thus be passed on to their affected child(ren). As such, mosaicism may be masking the true inheritance of disease and creating an overestimation of the *de novo* occurrence of WS. To investigate this possibility, we could use deep-coverage next-generation sequencing panels, similar to those currently used for testing tumour samples or affected tissues from other known somatic disorders; these panels are also likely to be more cost-effective for screening all known overgrowth genes in newly-recruited patients, when compared to Sanger sequencing.[3,340] High-throughput panels would not be as comprehensive as WES, as they are limited to sequencing a finite number of genes, but would provide sufficient coverage for detecting low-frequency somatic mosaicism. At this time, it is not known which line of investigation will have a higher diagnostic rate because somatic mutations in *EZH2* causing WS have not yet been identified. Importantly, determining the actual frequency of *EZH2* mosaic alterations could provide new guidelines for counselling parents on recurrence risks for future pregnancies (although gonadal mosaicism would be difficult to rule out).

#### 5.3.2.2.2 Functional studies

Another technical limitation of this thesis work involved the functional assay described in Chapter 3. Indeed, we observed significant background radioactivity even when using a known inactive mutant, which made low radioactivity counts virtually indistinguishable from the background counts. Unfortunately, this limitation could not be addressed with the equipment at our disposal. Ideally, a high-throughput scintillation counter should be used to recognize small changes in radioactivity; such counters usually read 96-well plates, which would also allow for a higher number of replicates at saturating conditions of the other substrates, while varying

concentrations of methyl acceptor or donor, so that we can determine steady-state kinetic parameters for each reaction and make our results much more robust. Alternatively, a luminescence label could be used on the methyl groups, and the luminescence signal would be read using a similar high-throughput machine.

### 5.3.3 Impact in the field

#### 5.3.3.1 Putting an end to the "diagnostic odyssey"

The greatest impact of this work is on the patients and families participating in our study. We were able to provide conclusive molecular results to seventeen families, including the two patients with mutations in *EED*, a gene that had not been previously associated with overgrowth. For these families, our results put an end to their emotionally demanding "diagnostic odyssey".[3,292,342] Indeed, it is important to remember that 25% of patients with rare diseases spend 5-30 years chasing a definitive diagnosis, and 50% never receive one.[3] Therefore, our results brought much reassurance to these families, and allowed them to connect with other affected families and support groups.[342] In some occasions, after being validated by a clinically-certified laboratory, this diagnosis will alter the care of the patients and help determine appropriate screening protocols for potential comorbidities such as cancer.[342] In addition, it will also reduce the number of unnecessary invasive procedures carried out to characterize the disease,[3] and halt inadequate treatments,[3] altogether reducing the costs of healthcare for both families and healthcare providers.[291] Lastly, by discovering a novel cause of overgrowth and sharing our findings through scientific publications, we have already contributed to the diagnosis of patients in other centers across the world (based on personal communications), and will likely see *EED* sequencing become available as a clinical test for overgrowth in the near future.

#### 5.3.3.2 Treatment of Weaver syndrome

As mentioned in Chapter 3, EZH2 inhibitors are currently being developed,[260–263,301,439] with two compounds in early stage clinical trials.[440,441] Although these were developed to treat cancer, they also represent an attractive treatment option for individuals with constitutional *EZH2* mutations, mainly WS patients. It is clear that developmental malformations cannot be "undone", so the dysmorphism and congenital abnormalities of various organs would remain. However, the intellectual disability could potentially be attenuated if the correct treatment were provided at an

early age, particularly in patients with no serious brain abnormalities. Indeed, EZH2 has been shown to be involved in cortical development, [289] in neuronal migration and connectivity,[290] and more importantly in neural differentiation which persists after birth.[442] Importantly, our data suggest that these inhibitors would not be beneficial to WS patients, and could even be harmful to them (see Chapter 3). Our work represents the first step towards advancing the care of WS patients with confirmed constitutional mutations in *EZH2*; further investigations (such as measuring EZH2 levels in WS patients, as mentioned earlier) will be required to progress these findings. In addition, the study of cancer development in patients with constitutional mutations in *EZH2* may also inform us on the role of EZH2 in cancer pathogenesis in general, meaning that insights gained from studying WS patients may actually be applicable to the wider population.

## 5.4 Future directions

### 5.4.1 Establishing a longitudinal follow-up of patients

As a follow-up on the work presented in Chapter 2, a definitive database of Weaver and Weaver-like patients should be established, ideally allowing for longitudinal studies that could shed a light on possible long-term health complications. For example, following WS patients over many years may give us a better estimate of the total lifetime cancer risk and help us determine the appropriate screening protocol (and adjust the surveillance plan suggested in Chapter 2, section 2.4.1.1). Gathering such information will be most valuable for parents who want to know the health prognostics of their child and plan appropriate care for the future.

This database should be as comprehensive as possible, as determined during this study, but standardized. The use of HPO terms employed in PhenoTips may assist in this standardization.[401] However, the database must also allow for personal comments, since unique characteristics are important to report. In addition, phenotypic information should be accompanied by age of assessment, so that progress over the years may be followed. Finally, each entry should provide detail regarding molecular testing results and variant interpretation rationale (for example whether pathogenicity was determined based on population genetics information alone or in combination with functional studies, and/or linking this entry to other affected individuals). Ideally, this database would have different interfaces for families and physicians/researchers, so

that families may access important information about health outcomes without depending on their physicians.

### 5.4.2 Determining the mechanism of disease for *EZH2* and *EED* mutations in overgrowth syndromes

In Chapter 3 we presented some *in vitro* functional work that suggested that WS may be caused by an impairment in PRC2-mediated H3K27 methylation. However, our results also showed that this assay is not reliable to predict the pathogenicity of novel variants in *EZH2*, because one mutant predicted to have a function equivalent to wild-type (D185H) also showed impaired activity *in vitro*. Therefore, additional functional investigations should be pursued.

As mentioned earlier (see Chapter 3, section 3.4.1), the most reliable source of material for investigation would be patient-derived samples, rather than protein mutants synthesized *in vitro*. Ideally, we would generate lymphoblastoid cell lines (LCLs) from patients' blood and, in comparison with LCLs derived from healthy controls, measure absolute levels of EZH2/EED expression and H3K27 methylation, as well as the antagonistic histone marks H3K27ac and H3K4me3. Such experiments would be carried out under the assumption that functional changes are not limited to specific tissues or developmental stages, and would be measurable in these patient-derived samples. We expect loss-of-function mutations to reduce or abolish H3K27me3 at PRC2-regulated sites, thereby derepressing target genes and/or allowing for activating marks to be deposited; in contrast, gain-of-function mutations should increase H3K27me3 at these promoters, thereby repressing these targets. Importantly, we may not detect any differences in global levels of H3K27me3, although such a finding would not necessarily constitute a negative result; instead, it could reflect lack of sensitivity from the assay used to measure mild changes in methylation levels, or it could illustrate that global levels of H3K27me3 remain the same while a re-distribution of repressive marks occurs across the genome. To investigate this possibility, we would need to carry out H3K27me1/2/3 ChIP-seq on these samples. This strategy may be used both for patients with WS and patients with *EED*-related overgrowth, since both proteins are required for PRC2-mediated H3K27 methylation; these experiments may also help us determine whether these mutants are mechanistically distinct, or whether they are likely to reflect a single disorder instead. Overall, these results may help understand the disease phenotypes, and could potentially shed a light on novel therapeutic avenues.

In using such a high-throughput line of investigation, several limitations remain. First of all, we may not be able to collect truly fresh samples from patients due to geographic distance (as mentioned earlier). In such case, we could consider the use of CRISPR-Cas9 technology to generate cell lines with the same mutations.[443,444] This technology has already been successfully employed to inactivate *EZH2* in a human cancer cell line,[445] but our model would require replacement of only one wild-type allele with the mutant allele, which is much more complex. In addition, this method is known to have many off-target effects which may interfere with our results (meaning that we wouldn't be able to determine if observed changes are actually due to our mutation or secondary hits).[446,447] In addition, we are assuming that H3K27 is the main substrate of EZH2-mediated methylation; another substrate may be a more critical determinant of the overgrowth and dysmorphic phenotypes observed in these patients. For a more complete functional assessment, we should also investigate the effects of these mutations on EZH2's ability to methylate its other known substrates such as JARID2-K116,[331] GATA4-K299,[332] STAT3-K180,[330] RORα-K38,[333] and H1b-K26.[201] Finally, a better characterization of the PRC2-independent functions of EZH2 and EED may also be necessary.

### 5.4.3 Diagnosing the rest of the cohort

In Chapter 2 we described work that identified pathogenic mutations in *EZH2* and *NSD1* in a subset of patients from our cohort, and in Chapter 4 we described work carried out to investigate new causes of disease in a subset of the remaining undiagnosed patients; continuing efforts should be made to diagnose the rest of the cohort.

#### 5.4.3.1 Definitively classifying variants of uncertain significance

First and foremost, investigations should be carried out in an attempt to conclusively classify the VOUS found in *EZH2* and *NSD1*. To definitively classify the D185H variant, we will require further functional work, as described above. For potential splice variants, we can collect patient RNA (with specialized saliva kits) and carry out RT-PCR to determine whether splicing is altered. We may also consider RNA-seq studies to rule out non-canonical splicing errors due to other variants in these genes.[296]

In addition, for these variants as well as other variants where evidence from population genetics is not sufficient to establish pathogenicity, there is an emerging new technique that may

assist in variant interpretation: we call it "epiprofiling". Choufani *et al*.[448] have recently shown that *NSD1* positive samples from Sotos syndrome patients share a common genome-wide DNA methylation signature, or "epiprofile". This signature represents differential methylation status at 7,085 CpG sites distributed across the genome, and is distinguishable from the signature generated at these CpG sites by samples from healthy controls or *EZH2* positive samples from WS patients. When the authors attempted to predict the *NSD1* mutational status of a new set of samples, blinded to their corresponding clinical diagnosis, and based solely on epigenetic clustering using this specific DNA methylation profile, they achieved 100% concordance. Furthermore, the epiprofile is so robust that it is retained across samples despite variables such as sex, age, and cell-type composition (for blood-derived samples), with preliminary data suggesting that the DNA methylation signature can even be recognized across different tissues, meaning that the changes in DNA methylation caused by alterations of NSD1 function far surpass those observed across different tissues at these specific CpG sites. Importantly, this epiprofile can only interrogate *NSD1* mutational status, and a different epiprofile will have to be used to interrogate *EZH2* mutational status. Despite this compelling evidence, the main limitation of this assay is that DNA methylation is not a direct output of NSD1 or EZH2 function because both proteins are involved in the methylation of lysine residues on histone tails. Indeed, although EZH2 has been shown to interact directly with DNA methyltransferases (DNMTs) and is thought to play a role in recruiting these DNMTs to establish local DNA methylation in order to maintain gene silencing (because DNA methylation is believed to be a more stable silencing mark than H3K27 methylation),[207] the direct mechanistic link between histone methylation and DNA methylation remains poorly understood.[449] Therefore, this "epiprofiling" method should be used as an additional line of evidence in conjunction with clinical data and other prediction methods described earlier, and should not be used on its own to determine the pathogenicity of *NSD1* or *EZH2* variants (or variants in any other genes subsequently found to have a similarly unique "epiprofile", possibly other chromatin regulators known to cause overgrowth such as *DNMT3A* and *SETD2*).

#### 5.4.3.2 Investigating variants in novel candidate genes

Next, we should focus on patients who have already had exome sequencing done and in whom we identified a plausible causative variant in a gene that has not yet been associated with

disease. Data from these patients should be deposited into secure databases such as Phenotips, and shared through PhenomeCentral/Matchmaker Exchange (as described earlier)[401,402] in an attempt to find other cases across the world with significant phenotypic overlap and common candidate variants. In the best case scenario, we will be able to find additional cases to support our finding; if not, we will have to carry out functional work, as discussed in Chapter 4, section 4.5.1.3.

#### 5.4.3.3 Carrying out more sequencing

Exome data of undiagnosed patients in whom we did not identify any candidate variants within the time-frame of this thesis should be re-analyzed routinely (every 6-12 months), to incorporate novel gene discoveries in the field; newly identified variants with conclusive classification should then be deposited into public variant databases in order to assist future variant interpretation. Other causal hypotheses such as homozygous recessive disorders or two gene disorders should also be explored. Furthermore, exome sequencing should be considered for patients who have not yet been investigated in this gene-agnostic manner, and their data should also be re-analyzed routinely and shared on PhenomeCentral (if needed). Carrying out sequencing in additional patients may also help us identify cases with overlapping candidate genes, which would in turn help us establish pathogenicity of novel genes (as described previously).

Additionally, expansion of screening to investigate regulatory regions (via Sanger or whole genome sequencing) may be contemplated. However, these non-coding regions are very repetitive, which makes them hard to amplify and sequence, and it would be challenging to definitively link each regulatory region to the specific gene of interest. Moreover, screening of the *NSD1* promoter region did not identify any alterations in patients described as having typical Sotos syndrome features and no mutations or microdeletions in the *NSD1* coding region; to note, this was a small study of only 18 patients, and there was no information provided regarding further testing carried out to investigate other causes of disease.[450] Together, this evidence suggests that investigating regulatory regions of known overgrowth genes may not be the most time- and cost-efficient strategy to seek a diagnosis for the remaining undiagnosed patients, and that, at least in the near future, we should still focus on interrogating coding regions to identify novel disease-causing variants.

## References


1. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* **14,** 681–691 (2013).
2. Boycott, K. *et al.* The clinical application of genome-wide sequencing for monogenic diseases in Canada: Position Statement of the Canadian College of Medical Geneticists. *J. Med. Genet.* **52,** 431–437 (2015).
3. Sawyer, S. L. *et al.* Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. *Clin. Genet.* **89,** 275–284 (2016).
4. Saxena, A. & Sampson, J. R. Phenotypes associated with inherited and developmental somatic mutations in genes encoding mTOR pathway components. *Semin. Cell Dev. Biol.* **36,** 140–146 (2014).
5. Keppler-Noreuil, K. M. *et al.* PIK3CA-related overgrowth spectrum (PROS): Diagnostic and testing eligibility criteria, differential diagnosis, and evaluation. *Am. J. Med. Genet.* **167A,** 287–295 (2015).
6. Erickson, R. P. Recent advances in the study of somatic mosaicism and diseases other than cancer. *Curr. Opin. Genet. Dev.* **26,** 73–78 (2014).
7. Yachelevich, N. Generalized overgrowth syndromes with prenatal onset. *Curr. Probl. Pediatr. Adolesc. Health Care* **45,** 97–111 (2015).
8. Malan, V. *et al.* Array-based comparative genomic hybridization identifies a high frequency of copy number variations in patients with syndromic overgrowth. *Eur. J. Hum. Genet.* **18,** 227–232 (2010).
9. Baujat, G. *et al.* Clinical and molecular overlap in overgrowth syndromes. *Am. J. Med. Genet.* **137C,** 4–11 (2005).
10. Weaver, D. D., Graham, C. B., Thomas, I. T. & Smith, D. W. A new overgrowth syndrome with accelerated skeletal maturation, unusual facies, and camptodactyly. *J. Pediatr.* **84,** 547–552 (1974).
11. Gemme, G., Bonioli, E., Ruffa, G. & Lagorio, V. The Weaver-Smith syndrome. *J. Pediatr.* **97,** 962–964 (1980).
12. Weisswichert, P. H., Knapp, G. & Willich, E. Accelerated Bone Maturation Syndrome of

the Weaver Type. *Eur. J. Pediatr.* **137,** 329–333 (1981).

13. Majewski, F., Ranke, M., Kemperdick, H. & Schmidt, E. The Weaver syndrome: a rare type of primordial overgrowth. *Eur. J. Pediatr.* **137,** 277–282 (1981).
14. Meinecke, P., Schaefer, E. & Engelbrccht, R. The Weaver syndrome in a girl. *Eur. J. Pediatr.* **141,** 58–59 (1983).
15. Roussounis, S. H. & Crawford, M. J. Siblings with Weaver syndrome. *J. Pediatr.* **102,** 595–597 (1983).
16. Tsukahara, M., Tanaka, S. & Kajii, T. A Weaver-like syndrome in a Japanese boy. *Clin. Genet.* **25,** 73–78 (1984).
17. Dawood, A. A., Machado, G. T. & Winship, W. S. Weaver's syndrome - primordial excessive growth velocity. A case report. *South African Med. J.* **67,** 646–648 (1985).
18. Farrell, S. A. & Hughes, H. E. Weaver syndrome with pes cavus. *Am. J. Med. Genet.* **21,** 737–739 (1985).
19. Stoll, C., Talon, P., Mengus, L., Roth, M. P. & Dott, B. A Weaver-like syndrome with endocrinological abnormalities in a boy and his mother. *Clin. Genet.* **28,** 255–259 (1985).
20. Ardinger, H. H. *et al.* Further delineation of Weaver syndrome. *J. Pediatr.* **108,** 228–235 (1986).
21. Thompson, E. M., Hill, S., Leonard, J. V & Pembrey, M. E. A girl with the Weaver syndrome. *J. Med. Genet.* **24,** 232–234 (1987).
22. Fretzayas, A., Papanicolaou, A., Tzanetakos, K., Theodoridis, C. & Karpathios, T. Retarded Skeletal Maturation in Weaver Syndrome. *Acta Paediatr. Scand.* **77,** 930–932 (1988).
23. Greenberg, F., Wasiewski, W. & McCabe, E. R. Weaver syndrome: the changing phenotype in an adult. *Am. J. Med. Genet.* **33,** 127–129 (1989).
24. Teebi, A. S., Sundareshan, T. S., Hammouri, M. Y., Al-Awadi, S. A. & Al-Saleh, Q. A. Teebi_1989_AmJMedGenet. *Am. J. Med. Genet.* **33,** 479–482 (1989).
25. Smyth, R. L., Gould, J. D. & Baraitser, M. A case of Marshall-Smith or Weaver syndrome. *J. R. Soc. Med.* **82,** 682–683 (1989).
26. Muhonen, M. G. & Menezes, A. H. Weaver syndrome and instability of the upper cervical spine. *J. Pediatr.* **116,** 596–599 (1990).
27. Trabelsi, M., Ben Hariz, M., Monastiri, K., Taktak, M. & Bennaceur, B. Weaver's

syndrome. Apropos of a new case. *Ann. Pediatr.* **37,** 327–330 (1990).

28. Kondo, I., Mori, Y. & Kuwajima, K. A Japanese male infant with the Weaver syndrome. *Jpn. J. Hum. Genet.* **35,** 257–262 (1990).
29. Kondo, I., Mori, Y. & Kuwajima, K. Weaver syndrome in two Japanese children. *Am. J. Med. Genet.* **41,** 221–224 (1991).
30. Ramos-Arroyo, M. A., Weaver, D. D. & Banks, E. R. Weaver Syndrome : A Case Without Early Overgrowth and Review of the Literature. *Pediatrics* **88,** 1106–1111 (1991).
31. Cole, T. R., Dennis, N. R. & Hughes, H. E. Weaver syndrome. *J. Med. Genet.* **29,** 332–337 (1992).
32. Dumic, M., Vukovic, J., Cvitkovic, M. & Medica, I. Twins and their mildly affected mother with Weaver syndrome. *Clin. Genet.* **44,** 338–340 (1993).
33. Scarano, G., Della Monica, M., Lonardo, F. & Neri, G. Novel findings in a patient with weaver or a weaver-like syndrome. *Am. J. Med. Genet.* **63,** 378–381 (1996).
34. Nishimura, G., Hasegawa, T. & Nagai, T. Propositus with Weaver syndrome and his mildly-affected mother: implication of nontraditional inheritance? *Am. J. Med. Genet.* **65,** 249–251 (1996).
35. Sanchez, O., Boufajreldin, S., Oranges, C., Orta, C. & Guerra, D. Weaver syndrome. 1st case reported in Venezuela. *Invest. Clin.* **38,** 9–24 (1997).
36. Fryer, A., Smith, C., Rosenbloom, L. & Cole, T. Autosomal dominant inheritance of Weaver syndrome. *J. Med. Genet.* **34,** 418–419 (1997).
37. Opitz, J. M., Weaver, D. W. & Reynolds, J. F. J. The syndromes of Sotos and Weaver: Reports and review. *Am. J. Med. Genet.* **79,** 294–304 (1998).
38. Proud, V. K., Braddock, S. R., Cook, L. & Weaver, D. D. Weaver syndrome: Autosomal dominant inheritance of the disorder. *Am. J. Med. Genet.* **79,** 305–310 (1998).
39. Sarigül, A., Yilmaz, M., Ateş, S. & Yurdakul, Y. A case with Weaver syndrome operated for congenital cardiac defect. *Pediatr. Cardiol.* **20,** 375–376 (1999).
40. Freeman, B., Hoon, A., Breiter, S. & Hamosh, A. Pachygyria in Weaver Syndrome. *Am. J. Med. Genet.* **86,** 395–397 (1999).
41. Derry, C., Temple, I. K. & Venkat-Raman, K. A probable case of familial Weaver syndrome associated with neoplasia. *J. Med. Genet.* **36,** 725–728 (1999).

42. Ozkan, B. & Bereket, A. Excessive growth in a girl with Weaver syndrome. *J. Pediatr. Endocrinol. Metab.* **13,** 1147–1153 (2000).
43. Kelly, T. E., Alford, B. A. & Abel, M. Cervical spine anomalies and tumors in Weaver syndrome. *Am. J. Med. Genet.* **95,** 492–495 (2000).
44. Huffman, C. *et al.* Weaver syndrome with neuroblastoma and cardiovascular anomalies. *Am. J. Med. Genet.* **99,** 252–255 (2001).
45. Voorhoeve, P. G., van Gils, J. F. & Jansen, M. The difficulty of height prediction in Weaver syndrome. *Clin. Dysmorphol.* **11,** 49–52 (2002).
46. Türkmen, S. *et al.* Mutations in NSD1 are responsible for Sotos syndrome, but are not a frequent finding in other overgrowth phenotypes. *Eur. J. Hum. Genet.* **11,** 858–865 (2003).
47. Miyoshi, Y. *et al.* Hormonal and genetical assessment of a Japanese girl with Weaver syndrome. *Clin. Pediatr. Endocrinol.* **13,** 17–23 (2004).
48. Crawford, M. W. & Rohan, D. The upper airway in Weaver syndrome. *Paediatr. Anaesth.* **15,** 893–896 (2005).
49. Iatrou, I. A., Schoinohoriti, O. K., Tzerbos, F. & Pasparakis, D. Treatment of macroglossia in a child with Weaver syndrome. *Int. J. Oral Maxillofac. Surg.* **37,** 961–965 (2008).
50. Coulter, D., Powell, C. M. & Gold, S. Weaver syndrome and neuroblastoma. *J. Pediatr. Hematol. Oncol.* **30,** 758–760 (2008).
51. Bansal, N. & Bansal, A. Weaver syndrome: A report of a rare genetic syndrome. *Indian J. Hum. Genet.* **15,** 36–37 (2009).
52. Basel-Vanagaite, L. Acute lymphoblastic leukemia in weaver syndrome. *Am. J. Med. Genet.* **152A,** 383–386 (2010).
53. Mikalef, P. *et al.* Weaver syndrome associated with bilateral congenital hip and unilateral subtalar dislocation. *Hippokratia* **14,** 212–214 (2010).
54. Miller, K., Abukabbos, H. & Mugayar, L. Oral, radiographical, and clinical findings in Weaver syndrome: A case report. *Spec. Care Dent.* **35,** 253–257 (2015).
55. Al-Salem, A., Alshammari, M. J., Hassan, H., Alazami, A. M. & Alkuraya, F. S. Weaver syndrome and defective cortical development: A rare association. *Am. J. Med. Genet.* **161A,** 225–227 (2013).

56. Gibson, W. T. *et al.* Mutations in EZH2 cause Weaver syndrome. *Am. J. Hum. Genet.* **90,** 110–118 (2012).
57. Tatton-Brown, K. *et al.* Germline mutations in the oncogene EZH2 cause Weaver syndrome and increased human height. *Oncotarget* **2,** 1127–1133 (2011).
58. Tatton-Brown, K. *et al.* Weaver syndrome and EZH2 mutations: Clarifying the clinical phenotype. *Am. J. Med. Genet.* **161A,** 2972–2980 (2013).
59. Cohen, A. S. A. *et al.* Weaver syndrome-associated EZH2 protein variants show impaired histone methyltransferase function in vitro. *Hum. Mutat.* **37,** 301–307 (2016).
60. Usemann, J., Ernst, T., Sch, V., Lehmberg, K. & Seeger, K. EZH2 Mutation in an Adolescent With Weaver Syndrome Developing Acute Myeloid Leukemia and Secondary Hemophagocytic Lymphohistiocytosis. *Am. J. Med. Genet.* **170,** 1274–1277 (2016).
61. Tatton-Brown, K. & Rahman, N. The NSD1 and EZH2 overgrowth genes, similarities and differences. *Am. J. Med. Genet.* **163C,** 86–91 (2013).
62. Jones, K. L., Jones, M. C. & Campo, M. del. *Smith's recognizable patterns of human malformation*. (2013).
63. Sotos, J. F., Dodge, P. R., Muirhead, D., Crawford, J. D. & Talbot, N. B. Cerebral Gigantism in Childhood - A Syndrome of Excessively Rapid Growth with Acromegalic Features and a Nonprogressive Neurologic Disorder. *N. Engl. J. Med.* **271,** 109–116 (1964).
64. Hook, E. B. & Reynolds, J. W. Cerebral gigantism: endocrinological and clinical observations of six patients including a congenital giant, concordant monozygotic twins, and a child who acheived adult gigantic size. *J. Pediatr.* **70,** 900–914 (1967).
65. Yeh, H., Price, R. L. & Lonsdale, D. Cerebral Gigantism (Sotos' Syndrome) and Cataracts. *J. Pediatr. Ophthalmol. Strabismus* **15,** 231–232 (1978).
66. Ferrier, P. E., de Meuron, G., Korol, S. & Hauser, H. Cerebral gigantism (Sotos syndrome) with juvenile macular degeneration. *Helv. Paediatr. Acta* **35,** 97–102 (1980).
67. Livieri, C., Gelmi, C. G., Lorini, R. & Gasparoni, A. Retinal degeneration in Sotos' syndrome. *Helv. Paediatr. Acta* **37,** 93–94 (1982).
68. Maino, D. M., Kofman, J., Flynn, M. F. & Lai, L. Ocular manifestations of Sotos syndrome. *J. Am. Optom. Assoc.* **65,** 339–346 (1994).
69. Kaneko, H. *et al.* Congenital heart defects in Sotos sequence. *Am. J. Med. Genet.* **26,** 569–

576 (1987).

70. Noreau, D. R., Al-Ata, J., Jutras, L. & Teebi, A. S. Congenital heart defects in Sotos syndrome. *Am. J. Med. Genet.* **79,** 327–328 (1998).
71. Miyamoto, T., Kitahori, K., Miyaji, K., Nagata, N. & Yasui, S. Total cavopulmonary connection in a bedridden patient with Sotos syndrome. *Asian Cardiovasc. Thorac. Ann.* **11,** 342–343 (2003).
72. Saccucci, P. *et al.* Isolated left ventricular noncompaction in a case of sotos syndrome: a casual or causal link? *Cardiol. Res. Pract.* **2011,** 824095 (2011).
73. Sotos, J. F., Romshe, C. A. & Cutler, E. A. Cerebral gigantism and primary hypothyroidism: pleiotropy or incidental concurrence. *Am. J. Med. Genet.* **2,** 201–205 (1978).
74. Ranke, M. B. & Bierich, J. R. Cerebral gigantism of hypothalamic origin. *Eur. J. Pediatr.* **140,** 109–111 (1983).
75. Bale, A. E., Drum, M. A., Parry, D. M. & Mulvihill, J. J. Familial Sotos syndrome (cerebral gigantism): craniofacial and psychological characteristics. *Am. J. Med. Genet.* **20,** 613–624 (1985).
76. De Boer, L. *et al.* Mutations in the NSD1 gene in patients with Sotos syndrome associate with endocrine and paracrine alterations in the IGF system. *Eur. J. Endocrinol.* **151,** 333–341 (2004).
77. de Boer, L. *et al.* Plasma insulin-like growth factors (IGFs), IGF-Binding proteins (IGFBPs), acid-labile subunit (ALS) and IGFBP-3 proteolysis in individuals with clinical characteristics of Sotos syndrome. *J. Pediatr. Endocrinol. Metab.* **17,** 615–627 (2004).
78. Zechner, U. *et al.* Familial Sotos syndrome caused by a novel missense mutation, C2175S, in NSD1 and associated with normal intelligence, insulin dependent diabetes, bronchial asthma, and lipedema. *Eur. J. Med. Genet.* **52,** 306–310 (2009).
79. Matsuo, T. *et al.* Hyperinsulinemic hypoglycemia of infancy in Sotos syndrome. *Am. J. Med. Genet.* **161A,** 34–37 (2013).
80. Wejaphikul, K. *et al.* Hypoparathyroidism in a 3-year-old Korean boy with Sotos syndrome and a novel mutation in NSD1. *Ann. Clin. Lab. Sci.* **45,** 215–218 (2015).
81. Nakamura, Y. *et al.* A case with neonatal hyperinsulinemic hypoglycemia: It is a characteristic complication of Sotos syndrome. *Am. J. Med. Genet.* **167A,** 1171–1174

(2015).

82. Haga, N., Nakamura, S., Shimode, M., Yanagisako, Y. & Iwaya, T. Scoliosis in cerebral gigantism, Sotos syndrome. A case report. *Spine (Phila. Pa. 1976)*. **21,** 1699–1702 (1996).
83. Carlo, W. & Dormans, J. P. Cervical instability in Sotos syndrome: a case report. *Spine (Phila. Pa. 1976)*. **29,** E153-156 (2004).
84. Tsirikos, A. I., Demosthenous, N. & McMaster, M. J. Spinal deformity in patients with Sotos syndrome (cerebral gigantism). *J. Spinal Disord. Tech.* **22,** 149–153 (2009).
85. Corrado, R. *et al.* Sotos syndrome and scoliosis surgical treatment: a 10-year follow-up. *Eur. Spine J.* **20 Suppl 2,** S271-277 (2011).
86. Compton, M. T., Celentana, M., Price, B. & Furman, A. C. A case of Sotos syndrome (cerebral gigantism) and psychosis. *Psychopathology* **37,** 190–193 (2004).
87. Nicita, F. *et al.* Seizures and epilepsy in Sotos syndrome: analysis of 19 Caucasian patients with long-term follow-up. *Epilepsia* **53,** e102-105 (2012).
88. Sheth, K. *et al.* The behavioral characteristics of Sotos syndrome. *Am. J. Med. Genet.* **167A,** 2945–2956 (2015).
89. Lane, C., Milne, E. & Freeth, M. Cognition and behaviour in Sotos syndrome: A systematic review. *PLoS One* **11,** 1–21 (2016).
90. Callanan, A. P., Anand, P. & Sheehy, E. C. Sotos syndrome with hypodontia. *Int. J. Paediatr. Dent.* **16,** 143–146 (2006).
91. Kotilainen, J., Pohjola, P., Pirinen, S., Arte, S. & Nieminen, P. Premolar hypodontia is a common feature in Sotos syndrome with a mutation in the NSD1 gene. *Am. J. Med. Genet.* **149A,** 2409–2414 (2009).
92. Hirai, N., Matsune, K. & Ohashi, H. Craniofacial and oral features of Sotos syndrome: Differences in patients with submicroscopic deletion and mutation of NSD1 gene. *Am. J. Med. Genet.* **155A,** 2933–2939 (2011).
93. Takano, M., Kasahara, K., Ogawa, C., Katada, H. & Sueishi, K. A case of Sotos syndrome treated with distraction osteogenesis in maxilla and mandible. *Bull. Tokyo Dent. Coll.* **53,** 75–82 (2012).
94. Hood, R. L. *et al.* Severe connective tissue laxity including aortic dilatation in Sotos syndrome. *Am. J. Med. Genet.* **170,** 531–535 (2016).
95. Cole, T. R. & Hughes, H. E. Sotos syndrome. *J. Med. Genet.* **27,** 571–576 (1990).

96. Allanson, J. E. & Cole, T. R. Sotos syndrome: evolution of facial phenotype subjective and objective assessment. *Am. J. Med. Genet.* **65,** 13–20 (1996).
97. Fickie, M. R. *et al.* Adults with Sotos syndrome: Review of 21 adults with molecularly confirmed NSD1 alterations, including a detailed case report of the oldest person. *Am. J. Med. Genet.* **155A,** 2105–2111 (2011).
98. Sugarman, G. I., Heuser, E. T. & Reed, W. B. A case of cerebral gigantism and hepatocarcinoma. *Am. J. Dis. Child.* **131,** 631–633 (1977).
99. Schrander-Stumpel, C. T., Fryns, J. P. & Hamers, G. G. Sotos syndrome and de novo balanced autosomal translocation (t(3;6)(p21;p21)). *Clin. Genet.* **37,** 226–229 (1990).
100. Cole, T. R., Hughes, H. E., Jeffreys, M. J., Williams, G. T. & Arnold, M. M. Small cell lung carcinoma in a patient with Sotos syndrome: are genes at 3p21 involved in both conditions? *J. Med. Genet.* **29,** 338–341 (1992).
101. Corsello, G. *et al.* Lymphoproliferative disorders in Sotos syndrome: observation of two cases. *Am. J. Med. Genet.* **64,** 588–593 (1996).
102. Cohen, M. M. Tumors and nontumors in Sotos syndrome. *Am. J. Med. Genet.* **84,** 173–175 (1999).
103. Le Marec, B., Pasquier, L., Dugast, C., Gosselin, M. & Odent, S. Gastric carcinoma in Sotos syndrome (cerebral gigantism). *Ann. Génétique* **42,** 113–116 (1999).
104. Leonard, N. J., Cole, T., Bhargava, R., Honoré, L. H. & Watt, J. Sacrococcygeal teratoma in two cases of Sotos syndrome. *Am. J. Med. Genet.* **95,** 182–184 (2000).
105. Jin, Y. *et al.* Sacrococcygeal germ cell tumor and spinal deformity in association with Sotos syndrome. *Med. Pediatr. Oncol.* **38,** 133–134 (2002).
106. Chen, C.-P. *et al.* Bilateral calcified ovarian fibromas in a patient with Sotos syndrome. *Fertil. Steril.* **77,** 1285–1287 (2002).
107. Al-Mulla, N., Belgaumi, A. F. & Teebi, A. Cancer in Sotos syndrome: report of a patient with acute myelocytic leukemia and review of the literature. *J. Pediatr. Hematol. Oncol.* **26,** 204–208 (2004).
108. Visser, R. *et al.* Identification of a 3.0-kb Major Recombination Hotspot in Patients with Sotos Syndrome Who Carry a Common 1.9-Mb Microdeletion. *Am. J. Hum. Genet.* **76,** 52–67 (2005).
109. Ruiz del Río, N., Abelairas Gómez, J. M., Peralta Calvo, J. M. & Miranda Lloret, P.

Atypical retinoblastoma in Sotos syndrome (cerebral gigantism). *Arch. Ophthalmol.* **125,** 578–580 (2007).

110. Martínez-Glez, V. & Lapunzina, P. Sotos syndrome is associated with leukemia/lymphoma. *Am. J. Med. Genet.* **143A,** 1244–1245 (2007).
111. Kato, M. *et al.* Hepatoblastoma in a Patient with Sotos Syndrome. *J. Pediatr.* **155,** 937–939 (2009).
112. Kulkarni, K., Stobart, K. & Noga, M. A case of Sotos syndrome with neuroblastoma. *J. Pediatr. Hematol. Oncol.* **35,** 238–239 (2013).
113. Beurdeley, M. *et al.* Ovarian Fibromatosis and Sotos Syndrome with a New Genetic Mutation. *J. Pediatr. Adolesc. Gynecol.* **26,** e39–e41 (2013).
114. Theodoulou, E., Baborie, A. & Jenkinson, M. D. Low grade glioma in an adult patient with Sotos syndrome. *J. Clin. Neurosci.* **22,** 413–415 (2015).
115. Nevo, S., Zeltzer, M., Benderly, A. & Levy, J. Evidence for autosomal recessive inheritance in cerebral gigantism. *J. Med. Genet.* **11,** 158–165 (1974).
116. Hansen, F. J. & Friis, B. Familial occurrence of cerebral gigantism, Sotos' syndrome. *Acta Paediatr. Scand.* **65,** 387–389 (1976).
117. Zonana, J., Rimoin, D. L. & Fisher, D. A. Cerebral gigantism - apparent dominant inheritance. *Birth Defects Orig. Artic. Ser.* **12,** 63–69 (1976).
118. Zonana, J. *et al.* Dominant inheritance of cerebral gigantism. *J. Pediatr.* **91,** 251–256 (1977).
119. Boman, H. & Nilsson, D. Sotos syndrome in two brothers. *Clin. Genet.* **18,** 421–427 (1980).
120. Winship, I. M. Sotos syndrome - autosomal dominant inheritance substantiated. *Clin. Genet.* **28,** 243–246 (1985).
121. Brown, W. T. *et al.* Identical twins discordant for Sotos syndrome. *Am. J. Med. Genet.* **79,** 329–333 (1998).
122. Chen, C.-P. *et al.* Perinatal imaging findings of inherited Sotos syndrome. *Prenat. Diagn.* **22,** 887–892 (2002).
123. Hoglund, P. *et al.* Familial Sotos syndrome is caused by a novel 1 bp deletion of the NSD1 gene. *J. Med. Genet.* **40,** 51–54 (2003).
124. van Haelst, M. M. *et al.* Familial gigantism caused by an NSD1 mutation. *Am. J. Med.*

*Genet.* **139A,** 40–44 (2005).

125. Tei, S., Tsuneishi, S. & Matsuo, M. The first Japanese familial Sotos syndrome with a novel mutation of the NSD1 gene. *Kobe J. Med. Sci.* **52,** 1–8 (2006).
126. Donnelly, D. E., Turnpenny, P. & McConnell, V. P. M. Phenotypic variability in a three-generation Northern Irish family with Sotos syndrome. *Clin. Dysmorphol.* **20,** 175–181 (2011).
127. Park, S. H., Lee, J. E., Sohn, Y. B. & Ko, J. M. First identified Korean family with Sotos syndrome caused by a novel intragenic mutation in NSD1. *Ann. Clin. Lab. Sci.* **44,** 228–231 (2014).
128. Kurotaki, N. *et al.* Haploinsufficiency of NSD1 causes Sotos syndrome. *Nat. Genet.* **30,** 365–366 (2002).
129. Kamimura, J. *et al.* Identification of eight novel NSD1 mutations in Sotos syndrome. *J. Med. Genet.* **40,** e126 (2003).
130. Rio, M. *et al.* Spectrum of NSD1 mutations in Sotos and Weaver syndromes. *J. Med. Genet.* **40,** 436–440 (2003).
131. Cecconi, M. *et al.* Mutation analysis of the NSD1 gene in a group of 59 patients with congenital overgrowth. *Am. J. Med. Genet.* **134A,** 247–253 (2005).
132. Waggoner, D. J. *et al.* NSD1 analysis for Sotos syndrome: insights and perspectives from the clinical laboratory. *Genet. Med.* **7,** 524–533 (2005).
133. Tatton-Brown, K. *et al.* Multiple mechanisms are implicated in the generation of 5q35 microdeletions in Sotos syndrome. *J. Med. Genet.* **42,** 307–313 (2005).
134. Tatton-Brown, K. *et al.* Genotype-phenotype associations in Sotos syndrome: an analysis of 266 individuals with NSD1 aberrations. *Am. J. Hum. Genet.* **77,** 193–204 (2005).
135. Malan, V. *et al.* Distinct effects of allelic NFIX mutations on nonsense-mediated mRNA decay engender either a Sotos-like or a Marshall-Smith syndrome. *Am. J. Hum. Genet.* **87,** 189–198 (2010).
136. Klaassens, M. *et al.* Malan syndrome: Sotos-like overgrowth with de novo NFIX sequence variants and deletions in six new patients and a review of the literature. *Eur. J. Hum. Genet.* **23,** 610–615 (2015).
137. Beckwith, J. B. Macroglossia, omphalocele, adrenal cytomegaly, gigantism and hyperplastic visceromegaly. *Birth Defects* **5,** 188–196 (1969).

138. Wiedemann, H. R. Complexe malformation familial avec hernie ombilicale et macroglossie - un 'syndrome nouveau'? *J. Génétique Hum.* **13,** 223–232 (1964).
139. Weksberg, R., Shuman, C. & Beckwith, J. B. Beckwith–Wiedemann syndrome. *Eur. J. Hum. Genet.* **18,** 8–14 (2010).
140. Choufani, S., Shuman, C. & Weksberg, R. Molecular findings in Beckwith-Wiedemann syndrome. *Am. J. Med. Genet.* **163C,** 131–140 (2013).
141. Maas, S. M. *et al.* Phenotype, cancer risk, and surveillance in Beckwith-Wiedemann syndrome depending on molecular genetic subgroups. *Am. J. Med. Genet.* **A,** (2016).
142. Mussa, A. *et al.* Cancer Risk in Beckwith-Wiedemann Syndrome: A Systematic Review and Meta-Analysis Outlining a Novel (Epi)Genotype Specific Histotype Targeted Screening Protocol. *J. Pediatr.* **176,** 142–149.e1 (2016).
143. Brioude, F. *et al.* Mutations of the Imprinted CDKN1C Gene as a Cause of the Overgrowth Beckwith-Wiedemann Syndrome: Clinical Spectrum and Functional Characterization. *Hum. Mutat.* **36,** 894–902 (2015).
144. Demars, J. & Gicquel, C. Epigenetic and genetic disturbance of the imprinted 11p15 region in Beckwith-Wiedemann and Silver-Russell syndromes. *Clin. Genet.* **81,** 350–361 (2012).
145. Milani, D., Pezzani, L., Tabano, S. & Miozzo, M. Beckwith-Wiedemann and IMAGe syndromes: Two very different diseases caused by mutations on the same gene. *Appl. Clin. Genet.* **7,** 169–175 (2014).
146. Pilia, G. *et al.* Mutations in GPC3, a glypican gene, cause the Simpson-Golabi-Behmel overgrowth syndrome. *Nat. Genet.* **12,** 241–247 (1996).
147. Neri, G., Marini, R., Cappa, M., Borrelli, P. & Opitz, J. M. Simpson-Golabi-Behmel syndrome: an X-linked encephalo-tropho-schisis syndrome. *Am. J. Med. Genet.* **30,** 287–299 (1988).
148. Simpson, J. L., Landey, S., New, M. & German, J. A previously unrecognized X-linked syndrome of dysmorphia. *Birth Defects* **XI,** 18–24 (1975).
149. Golabi, M. & Rosen, L. A new X-linked mental retardation-overgrowth syndrome. *Am. J. Med. Genet.* **17,** 345–358 (1984).
150. Behmel, A., Plöchl, E. & Rosenkranz, W. A new X-linked dysplasia gigantism syndrome: identical with the Simpson dysplasia syndrome? *Hum. Genet.* **67,** 409–413 (1984).

151. Tenorio, J. *et al.* Simpson-Golabi-Behmel syndrome types I and II. *Orphanet J. Rare Dis.* **9,** 138 (2014).
152. Budny, B. *et al.* A novel X-linked recessive mental retardation syndrome comprising macrocephaly and ciliary dysfunction is allelic to oral-facial-digital type I syndrome. *Hum. Genet.* **120,** 171–178 (2006).
153. Fauth, C. *et al.* A recurrent germline mutation in the PIGA gene causes Simpson-Golabi-Behmel syndrome type 2. *Am. J. Med. Genet.* **170A,** 392–402 (2016).
154. Rousseau, F., Rouillard, P., Morel, M. L., Khandjian, E. W. & Morgan, K. Prevalence of carriers of premutation-size alleles of the FMRI gene - and implications for the population genetics of the fragile X syndrome. *Am. J. Hum. Genet.* **57,** 1006–1018 (1995).
155. Suhl, J. A. & Warren, S. T. Single-Nucleotide Mutations in FMR1 Reveal Novel Functions and Regulatory Mechanisms of the Fragile X Syndrome Protein FMRP. *J. Exp. Neurosci.* **9,** 35–41 (2015).
156. Ramirez, F. & Dietz, H. C. Marfan syndrome: from molecular pathogenesis to clinical treatment. *Curr. Opin. Genet. Dev.* **17,** 252–258 (2007).
157. Pepe, G. *et al.* Marfan syndrome: current perspectives. *Appl. Clin. Genet.* **9,** 55–65 (2016).
158. Callewaert, B. L. *et al.* Comprehensive clinical and molecular assessment of 32 probands with congenital contractural arachnodactyly: report of 14 novel mutations and review of the literature. *Hum. Mutat.* **30,** 334–341 (2009).
159. Astuti, D. *et al.* Germline mutations in DIS3L2 cause the Perlman syndrome of overgrowth and Wilms tumor susceptibility. *Nat. Genet.* **44,** 277–284 (2012).
160. Tatton-Brown, K. *et al.* Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability. *Nat. Genet.* **46,** 385–388 (2014).
161. Luscan, A. *et al.* Mutations in SETD2 cause a novel overgrowth condition. *J. Med. Genet.* **51,** 512–517 (2014).
162. Lumish, H. S., Wynn, J., Devinsky, O. & Chung, W. K. Brief Report: SETD2 Mutation in a Child with Autism, Intellectual Disabilities and Epilepsy. *J. Autism Dev. Disord.* **45,** 3764–3770 (2015).
163. Tenorio, J. *et al.* A New Overgrowth Syndrome is due to Mutations in RNF125. *Hum. Mutat.* **35,** 1436–1441 (2014).
164. Takenouchi, T. *et al.* Novel overgrowth syndrome phenotype due to recurrent de novo

PDGFRB mutation. *J. Pediatr.* **166,** 483–486 (2015).

165. Ortega-Recalde, O. *et al.* Biallelic HERC1 mutations in a syndromic form of overgrowth and intellectual disability. *Clin. Genet.* **88,** e1-3 (2015).
166. Loveday, C. *et al.* Mutations in the PP2A regulatory subunit B family genes PPP2R5B, PPP2R5C and PPP2R5D cause human overgrowth. *Hum. Mol. Genet.* **24,** 4775–4779 (2015).
167. Thauvin-Robinet, C. *et al.* Homozygous FIBP nonsense variant responsible of syndromic overgrowth, with overgrowth, macrocephaly, retinal coloboma and learning disabilities. *Clin. Genet.* **89,** e1-4 (2016).
168. Akawi, N. *et al.* A recessive syndrome of intellectual disability, moderate overgrowth, and renal dysplasia predisposing to Wilms tumor is caused by a mutation in FIBP gene. *Am. J. Med. Genet.* **170,** 2111–2118 (2016).
169. Schäfgen, J. *et al.* De novo nonsense and frameshift variants of TCF20 in individuals with intellectual disability and postnatal overgrowth. *Eur. J. Hum. Genet.* (2016). doi:10.1038/ejhg.2016.90
170. Jenuwein, T. *et al.* Translating the histone code. *Science* **293,** 1074–1080 (2001).
171. Dupont, C., Armant, D. R. & Brenner, C. A. Epigenetics: definition, mechanisms and clinical perspective. *Semin. Reprod. Med.* **27,** 351–357 (2009).
172. Margueron, R. & Reinberg, D. Chromatin structure and the inheritance of epigenetic information. *Nat. Rev. Genet.* **11,** 285–296 (2010).
173. You, J. S. & Jones, P. A. Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell* **22,** 9–20 (2012).
174. Crea, F. & Crea, F. Histone code, human growth and cancer. *Oncotarget* **3,** 1–2 (2012).
175. Jones, R. S. & Gelbart, W. M. Genetic analysis of the enhancer of zeste locus and its role in gene regulation in Drosophila melanogaster. *Genetics* **126,** 185–199 (1990).
176. Paro, R. Mechanisms of heritable gene repression during development of Drosophila. *Curr. Opin. Cell Biol.* **5,** 999–1005 (1993).
177. LaJeunesse, D. & Shearn, A. E(z): a polycomb group gene or a trithorax group gene? *Development* **122,** 2189–2197 (1996).
178. Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P. & Reinberg, D. Histone methyltransferase activity associated with a human multiprotein complex

containing the enhancer of zeste protein. *Genes Dev.* **16,** 2893–2905 (2002).

179. Zuckerkandl, E. A possible role of 'inert' heterochromatin in cell differentiation. Action of and competition for 'locking' molecules. *Biochimie* **56,** 937–954 (1974).
180. Piunti, A. & Shilatifard, A. Epigenetic balance of gene expression by Polycomb and COMPASS families. *Science* **352,** aad9780 (2016).
181. Geisler, S. J. & Paro, R. Trithorax and Polycomb group-dependent regulation: a tale of opposing activities. *Development* **142,** 2876–2887 (2015).
182. Schuettengruber, B., Martinez, A.-M., Iovino, N. & Cavalli, G. Trithorax group proteins: switching genes on and keeping them active. *Nat. Rev. Mol. Cell Biol.* **12,** 799–814 (2011).
183. Simon, J. a & Kingston, R. E. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat. Rev. Mol. cell Biol.* **10,** 697–708 (2009).
184. Schwartz, Y. B. & Pirrotta, V. A new world of Polycombs: unexpected partnerships and emerging functions. *Nat. Rev. Genet.* **14,** 853–864 (2013).
185. Chen, H., Rossier, C. & Antonarakis, S. E. Cloning of a human homolog of the Drosophila enhancer of zeste gene (EZH2) that maps to chromosome 21q22.2. *Genomics* **38,** 30–37 (1996).
186. Laible, G. *et al.* Mammalian homologues of the Polycomb-group gene Enhancer of zeste mediate gene silencing in Drosophila heterochromatin and at S. cerevisiae telomeres. *EMBO J.* **16,** 3219–3232 (1997).
187. Jenuwein, T., Laible, G., Dorn, R. & Reuter, G. SET domain proteins modulate chromatin domains in eu- and heterochromatin. *Cell. Mol. Life Sci.* **54,** 80–93 (1998).
188. Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298,** 1039–1043 (2002).
189. Cao, R. & Zhang, Y. SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. *Mol. Cell* **15,** 57–67 (2004).
190. Sneeringer, C. J. *et al.* Coordinated activities of wild-type plus mutant EZH2 drive tumor-associated hypertrimethylation of lysine 27 on histone H3 (H3K27) in human B-cell lymphomas. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 20980–20985 (2010).
191. Kadoch, C., Copeland, R. A. & Keilhack, H. PRC2 and SWI/SNF Chromatin Remodeling Complexes in Health and Disease. *Biochemistry* **55,** 1600–1614 (2016).

192. Ferrari, K. J. *et al.* Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity. *Mol. Cell* **53,** 49–62 (2014).
193. Otte, A. P. & Kwaks, T. H. J. Gene repression by Polycomb group protein complexes: A distinct complex for every occasion? *Curr. Opin. Genet. Dev.* **13,** 448–454 (2003).
194. Alvarez-Venegas, R. & Avramova, Z. SET-domain proteins of the Su(var)3-9, E(z) and Trithorax families. *Gene* **285,** 25–37 (2002).
195. Sauvageau, M. & Sauvageau, G. Polycomb Group Proteins: Multi-Faceted Regulators of Somatic Stem Cells and Cancer. *Cell Stem Cell* **7,** 299–313 (2010).
196. Pasini, D. *et al.* JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. *Nature* **464,** 306–310 (2010).
197. Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. *Nat. Struct. Mol. Biol.* **20,** 1147–1155 (2013).
198. Swigut, T. & Wysocka, J. H3K27 demethylases, at long last. *Cell* **131,** 29–32 (2007).
199. Puri, D., Gala, H., Mishra, R. & Dhawan, J. High-wire act: The poised genome and cellular memory. *FEBS J.* **282,** 1675–1691 (2015).
200. Arcipowski, K. M., Martinez, C. A. & Ntziachristos, P. Histone demethylases in physiology and cancer: a tale of two enzymes, JMJD3 and UTX. *Curr. Opin. Genet. Dev.* **36,** 59–67 (2016).
201. Kuzmichev, A., Jenuwein, T., Tempst, P. & Reinberg, D. Different Ezh2-containing complexes target methylation of histone H1 or nucleosomal histone H3. *Mol. Cell* **14,** 183–193 (2004).
202. Martin, C., Cao, R. & Zhang, Y. Substrate preferences of the EZH2 histone methyltransferase complex. *J. Biol. Chem.* **281,** 8365–8370 (2006).
203. Margueron, R. *et al.* Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature* **461,** 762–767 (2009).
204. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17,** 487–500 (2016).
205. Jiao, L. & Liu, X. Structural basis of histone H3K27 trimethylation by an active polycomb repressive complex 2. *Science* **350,** aac4383 (2015).
206. Francis, N. J., Saurin, A. J., Shao, Z. & Kingston, R. E. Reconstitution of a functional core polycomb repressive complex. *Mol. Cell* **8,** 545–556 (2001).

207. Viré, E. *et al.* The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439,** 871–874 (2006).
208. Cha, T.-L. T.-L. *et al.* Akt-Mediated Phosphorylation of EZH2 Suppresses Methylation of Lysine 27 in Histone H3. *Science* **310,** 306–310 (2005).
209. Caretti, G., Palacios, D., Sartorelli, V. & Puri, P. L. Phosphoryl-EZH-ion. *Cell Stem Cell* **8,** 262–265 (2011).
210. Pengelly, A. R., Copur, Ö., Jäckle, H., Herzig, A. & Müller, J. A histone mutant reproduces the phenotype caused by loss of histone-modifying factor Polycomb. *Science* **339,** 698–699 (2013).
211. Lewis, P. W. *et al.* Inhibition of PRC2 Activity by a Gain-of-Function H3 Mutation Found in Pediatric Glioblastoma. *Science* **340,** 857–861 (2013).
212. Chan, K. *et al.* The histone H3.3K27M mutation in pediatric glioma reprograms H3K27 methylation and gene expression. *Genes Dev.* **27,** 985–990 (2013).
213. Pasini, D. *et al.* Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes. *Nucleic Acids Res.* **38,** 4958–4969 (2010).
214. Yuan, W. *et al.* H3K36 methylation antagonizes PRC2-mediated H3K27 methylation. *J. Biol. Chem.* **286,** 7983–9 (2011).
215. Cao, R. & Zhang, Y. The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. *Curr. Opin. Genet. Dev.* **14,** 155–164 (2004).
216. Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441,** 349–353 (2006).
217. Conway, E., Healy, E. & Bracken, A. P. PRC2 mediated H3K27 methylations in cellular identity and cancer. *Curr. Opin. Cell Biol.* **37,** 42–48 (2015).
218. Chen, Y.-H., Hung, M.-C. & Li, L.-Y. EZH2: a pivotal regulator in controlling cell differentiation. *Am. J. Transl. Res.* **4,** 364–375 (2012).
219. Voigt, P. *et al.* Asymmetrically Modified Nucleosomes. *Cell* **151,** 181–193 (2012).
220. Aldiri, I. & Vetter, M. L. PRC2 during vertebrate organogenesis: A complex in transition. *Dev. Biol.* **367,** 91–99 (2012).
221. O'Carroll, D. *et al.* The Polycomb-Group Gene Ezh2 Is Required for Early Mouse Development. *Mol. Cell. Biol.* **21,** 4330–4336 (2001).

222. Shumacher, A., Faust, C. & Magnuson, T. Positional cloning of a global regulator of anterior-posterior patterning in mice. *Nature* **383,** 250–253 (1996).
223. Pasini, D., Bracken, A. P., Jensen, M. R., Denchi, E. L. & Helin, K. Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. *EMBO J.* **23,** 4061–4071 (2004).
224. Richly, H., Aloia, L. & Di Croce, L. Roles of the Polycomb group proteins in stem cells and cancer. *Cell Death Dis.* **2,** e204 (2011).
225. Scelfo, A., Piunti, A. & Pasini, D. The controversial role of the Polycomb group proteins in transcription and cancer: how much do we not understand Polycomb proteins? *FEBS J.* **282,** 1703–1722 (2015).
226. Tavares, L. *et al.* RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3. *Cell* **148,** 664–678 (2012).
227. Voncken, J. W. *et al.* Rnf2 (Ring1b) deficiency causes gastrulation arrest and cell cycle inhibition. *Proc. Natl. Acad. Sci.* **100,** 2468–2473 (2003).
228. Endoh, M. *et al.* Histone H2A mono-ubiquitination is a crucial step to mediate PRC1-dependent repression of developmental genes to maintain ES cell identity. *PLoS Genet.* **8,** e1002774 (2012).
229. Lund, K., Adams, P. D. & Copland, M. EZH2 in normal and malignant hematopoiesis. *Leukemia* **28,** 44–49 (2014).
230. Ueda, T. *et al.* EED mutants impair polycomb repressive complex 2 in myelodysplastic syndrome and related neoplasms. *Leukemia* **26,** 2557–2560 (2012).
231. Score, J. *et al.* Inactivation of polycomb repressive complex 2 components in myeloproliferative and myelodysplastic/myeloproliferative neoplasms. *Blood* **119,** 1208–1213 (2012).
232. Ntziachristos, P. *et al.* Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. *Nat. Med.* **18,** 298–301 (2012).
233. Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481,** 157–163 (2012).
234. Laugesen, A. & Helin, K. Chromatin repressive complexes in stem cells, development, and cancer. *Cell Stem Cell* **14,** 735–751 (2014).
235. Plath, K. *et al.* Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300,**

131–135 (2003).

236. Silva, J. *et al.* Establishment of Histone H3 Methylation on the Inactive X Chromosome Requires Transient Recruitment of Eed-Enx1 Polycomb Group Complexes. *Dev. Cell* **4,** 481–495 (2003).
237. Terranova, R. *et al.* Polycomb Group Proteins Ezh2 and Rnf2 Direct Genomic Contraction and Imprinted Repression in Early Mouse Embryos. *Dev. Cell* **15,** 668–679 (2008).
238. Wu, H.-A. & Bernstein, E. Partners in Imprinting: Noncoding RNA and Polycomb Group Proteins. *Dev. Cell* **15,** 637–638 (2008).
239. Cardoso, C. *et al.* The human EZH2 gene: genomic organisation and revised mapping in 7q35 within the critical region for malignant myeloid disorders. *Eur. J. Hum. Genet.* **8,** 174–180 (2000).
240. Völkel, P., Dupret, B., Le Bourhis, X. & Angrand, P.-O. Diverse involvement of EZH2 in cancer epigenetics. *Am. J. Transl. Res.* **7,** 175–193 (2015).
241. Raaphorst, F. M. *et al.* Coexpression of BMI-1 and EZH2 Polycomb Group Genes in Reed-Sternberg Cells of Hodgkin's Disease. *Am. J. Pathol.* **157,** 709–715 (2000).
242. van Kemenade, F. J. *et al.* Coexpression of BMI-1 and EZH2 polycomb-group proteins is associated with cycling cells and degree of malignancy in B-cell non-Hodgkin lymphoma. *Blood* **97,** 3896–3901 (2001).
243. Varambally, S. *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419,** 624–629 (2002).
244. Kleer, C. G. *et al.* EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 11606–11611 (2003).
245. Weikert, S. *et al.* Expression levels of the EZH2 polycomb transcriptional repressor correlate with aggressiveness and invasive potential of bladder carcinomas. *Int. J. Mol. Med.* **16,** 349–353 (2005).
246. Lu, C. *et al.* Gene alterations identified by expression profiling in tumor-associated endothelial cells from invasive ovarian carcinoma. *Cancer Res.* **67,** 1757–1768 (2007).
247. Matsukawa, Y. *et al.* Expression of the enhancer of zeste homolog 2 is correlated with poor prognosis in human gastric cancer. *Cancer Sci.* **97,** 484–491 (2006).
248. Sudo, T. *et al.* Clinicopathological significance of EZH2 mRNA expression in patients

with hepatocellular carcinoma. *Br. J. Cancer* **92,** 1754–1758 (2005).

249. McHugh, J. B., Fullen, D. R., Ma, L., Kleer, C. G. & Su, L. D. Expression of polycomb group protein EZH2 in nevi and melanoma. *J. Cutan. Pathol.* **34,** 597–600 (2007).
250. Takawa, M. *et al.* Validation of the histone methyltransferase EZH2 as a therapeutic target for various types of human cancer and as a prognostic marker. *Cancer Sci.* **102,** 1298–1305 (2011).
251. Xu, K. *et al.* EZH2 Oncogenic Activity in Castration-Resistant Prostate Cancer Cells Is Polycomb-Independent. *Science* **338,** 1465–1469 (2012).
252. Cavalli, G. EZH2 Goes Solo. *Science* **338,** 1430–1431 (2012).
253. Gall Trošelj, K., Novak Kujundzic, R. & Ugarkovic, D. Polycomb repressive complex's evolutionary conserved function: the role of EZH2 status and cellular background. *Clin. Epigenetics* **8,** 55 (2016).
254. Morin, R. D. *et al.* Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat. Genet.* **42,** 181–185 (2010).
255. Lohr, J. G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 3879–3884 (2012).
256. Barsotti, A. M. *et al.* Epigenetic reprogramming by tumor-derived EZH2 gain-of-function mutations promotes aggressive 3D cell morphologies and enhances melanoma tumor growth. *Oncotarget* **6,** 2928–38 (2015).
257. Yap, D. B. *et al.* Somatic mutations at EZH2 Y641 act dominantly through a mechanism of selectively altered PRC2 catalytic activity, to increase H3K27 trimethylation. *Blood* **117,** 2451–2459 (2011).
258. McCabe, M. T. *et al.* Mutation of A677 in histone methyltransferase EZH2 in human B-cell lymphoma promotes hypertrimethylation of histone H3 on lysine 27 (H3K27). *Proc. Natl. Acad. Sci. U. S. A.* **109,** 2989–2994 (2012).
259. Majer, C. R. *et al.* A687V EZH2 is a gain-of-function mutation found in lymphoma patients. *FEBS Lett.* **586,** 3448–3451 (2012).
260. McCabe, M. T. *et al.* EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature* **492,** 108–112 (2012).
261. Qi, W. *et al.* Selective inhibition of Ezh2 by a small molecule inhibitor blocks tumor cells

proliferation. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 21360–21365 (2012).

262. Knutson, S. K. *et al.* A selective inhibitor of EZH2 blocks H3K27 methylation and kills mutant lymphoma cells. *Nat. Chem. Biol.* **8,** 890–896 (2012).
263. Knutson, S. K. *et al.* Durable tumor regression in genetically altered malignant rhabdoid tumors by inhibition of methyltransferase EZH2. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 7922–7927 (2013).
264. Ernst, T. *et al.* Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nat. Genet.* **42,** 722–726 (2010).
265. The Cancer Genome Atlas Research Network. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N. Engl. J. Med.* **368,** 2059–2074 (2013).
266. Chase, A. & Cross, N. C. P. Aberrations of EZH2 in cancer. *Clin. Cancer Res.* **17,** 2613–2618 (2011).
267. Martinez-Garcia, E. & Licht, J. D. Deregulation of H3K27 methylation in cancer. *Nat. Genet.* **42,** 100–101 (2010).
268. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
269. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
270. Jaju, R. J. *et al.* A novel gene, NSD1, is fused to NUP98 in the t(5;11)(q35;p15.5) in de novo childhood acute myeloid leukemia. *Blood* **98,** 1264–1267 (2001).
271. Kurotaki, N. *et al.* Molecular characterization of NSD1, a human homologue of the mouse Nsd1 gene. *Gene* **279,** 197–204 (2001).
272. Hollink, I. H. I. M. *et al.* NUP98/NSD1 characterizes a novel poor prognostic group in acute myeloid leukemia with a distinct HOX gene expression pattern. *Blood* **118,** 3645–3656 (2011).
273. Fasan, A. *et al.* A rare but specific subset of adult AML patients can be defined by the cytogenetically cryptic NUP98-NSD1 fusion gene. *Leukemia* **27,** 245–248 (2013).
274. Cerveira, N. *et al.* Frequency of NUP98-NSD1 fusion transcript in childhood acute myeloid leukaemia. *Leukemia* **17,** 2244–2247 (2003).
275. Rayasam, G. V. *et al.* NSD1 is essential for early post-implantation development and has a catalytically active SET domain. *EMBO J.* **22,** 3153–3163 (2003).

276. Li, Y. *et al.* The Target of the NSD Family of Histone Lysine Methyltransferases Depends on the Nature of the Substrate. *J. Biol. Chem.* **284,** 34283–34295 (2009).
277. Qiao, Q. *et al.* The structure of NSD1 reveals an autoregulatory mechanism underlying histone H3K36 methylation. *J. Biol. Chem.* **286,** 8361–8368 (2011).
278. Wang, G. G., Cai, L., Pasillas, M. P. & Kamps, M. P. NUP98-NSD1 links H3K36 methylation to Hox-A gene activation and leukaemogenesis. *Nat. Cell Biol.* **9,** 804–812 (2007).
279. Job, B. *et al.* Genomic aberrations in lung adenocarcinoma in never smokers. *PLoS One* **5,** e15145 (2010).
280. Bianco-Miotto, T. *et al.* Global levels of specific histone modifications and an epigenetic gene signature predict prostate cancer progression and development. *Cancer Epidemiol. Biomarkers Prev.* **19,** 2611–2622 (2010).
281. Berdasco, M. *et al.* Epigenetic inactivation of the Sotos overgrowth syndrome gene histone methyltransferase NSD1 in human neuroblastoma and glioma. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 21830–21835 (2009).
282. Fenton, T. R. *et al.* A systematic review and meta-analysis to revise the Fenton growth chart for preterm infants. *BMC Pediatr.* **13,** 59 (2013).
283. Douglas, J. *et al.* NSD1 mutations are the major cause of Sotos syndrome and occur in some cases of Weaver syndrome but are rare in other overgrowth phenotypes. *Am. J. Hum. Genet.* **72,** 132–143 (2003).
284. Johnston, J. J. & Biesecker, L. G. Databases of genomic variation and phenotypes: existing resources and future needs. *Hum. Mol. Genet.* **22,** R27–R31 (2013).
285. Xue, Y. *et al.* Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions, Mutation Databases, and Population-Scale Resequencing. *Am. J. Hum. Genet.* **91,** 1022–1032 (2012).
286. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498,** 220–223 (2013).
287. Bodian, D. L. *et al.* Germline variation in cancer-susceptibility genes in a healthy, ancestrally diverse cohort: implications for individual genome sequencing. *PLoS One* **9,** e94554 (2014).
288. Mirzaa, G. M. *et al.* Megalencephaly-capillary malformation (MCAP) and

megalencephaly-polydactyly-polymicrogyria-hydrocephalus (MPPH) syndromes: Two closely related disorders of brain overgrowth and abnormal brain and body morphogenesis. *Am. J. Med. Genet.* **158A,** 269–291 (2012).

289. Pereira, J. D. *et al.* Ezh2, the histone methyltransferase of PRC2, regulates the balance between self-renewal and differentiation in the cerebral cortex. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 15957–15962 (2010).
290. Di Meglio, T. *et al.* Ezh2 orchestrates topographic migration and connectivity of mouse precerebellar neurons. *Science* **339,** 204–207 (2013).
291. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97,** 199–215 (2015).
292. Tifft, C. J. & Adams, D. R. The National Institutes of Health undiagnosed diseases program. *Curr. Opin. Pediatr.* **26,** 626–633 (2014).
293. Mefford, H. C., Batshaw, M. L. & Hoffman, E. P. Genomics, intellectual disability, and autism. *N. Engl. J. Med.* **366,** 733–743 (2012).
294. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367,** 1921–1929 (2012).
295. Xiong, H. Y. *et al.* RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347,** 1254806 (2015).
296. Sibley, C. R., Blazquez, L. & Ule, J. Lessons from non-canonical splicing. *Nat. Rev. Genet.* **17,** 407–421 (2016).
297. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. *Cell* **156,** 1324–1335 (2014).
298. Zheng, S., Kim, H. & Verhaak, R. G. W. Silent Mutations Make Some Noise. *Cell* **156,** 1129–1131 (2014).
299. Yilmaz, R. *et al.* A recurrent synonymous *KAT6B* mutation causes Say-Barber-Biesecker/Young-Simpson syndrome by inducing aberrant splicing. *Am. J. Med. Genet.* **167A,** 3006–3010 (2015).
300. Ott, H. M. *et al.* A687V EZH2 is a driver of histone H3 lysine 27 (H3K27) hypertrimethylation. *Mol. Cancer Ther.* **13,** 3062–73 (2014).
301. Tan, J., Yan, Y., Wang, X., Jiang, Y. & Xu, H. E. EZH2: biology, disease, and structure-based drug discovery. *Acta Pharmacol. Sin.* **35,** 161–174 (2014).

302. Van Aller, G. S. *et al.* Long residence time inhibition of EZH2 in activated polycomb repressive complex 2. *ACS Chem. Biol.* **9,** 622–629 (2014).
303. Kim, W. *et al.* Targeted disruption of the EZH2–EED complex inhibits EZH2-dependent cancer. *Nat. Chem. Biol.* **9,** 643–650 (2013).
304. Makishima, H. *et al.* Novel homo- and hemizygous mutations in EZH2 in myeloid malignancies. *Leukemia* **24,** 1799–804 (2010).
305. Guglielmelli, P. *et al.* EZH2 mutational status predicts poor survival in myelofibrosis. *Blood* **118,** 5227–5234 (2011).
306. Schwarz, D. *et al.* Ezh2 is required for neural crest-derived cartilage and bone formation. *Development* **141,** 867–77 (2014).
307. Kudithipudi, S., Lungu, C., Rathert, P., Happel, N. & Jeltsch, A. Substrate specificity analysis and novel substrates of the protein lysine methyltransferase NSD1. *Chem. Biol.* **21,** 226–237 (2014).
308. Joshi, P. *et al.* Dominant alleles identify SET domain residues required for histone methyltransferase of Polycomb repressive complex 2. *J. Biol. Chem.* **283,** 27757–27766 (2008).
309. Tachibana, M., Sugimoto, K., Fukushima, T. & Shinkai, Y. SET Domain-containing Protein, G9a, is a Novel Lysine-preferring Mammalian Histone Methyltransferase with Hyperactivity and Specific Selectivity to Lysines 9 and 27 of Histone H3. *J. Biol. Chem.* **276,** 25309–25317 (2001).
310. Tachibana, M. *et al.* G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. *Genes Dev.* **16,** 1779–1791 (2002).
311. Bisswanger, H. Enzyme assays. *Perspect. Sci.* **1,** 41–55 (2014).
312. Grossmann, V. *et al.* EZH2 mutations and their association with PICALM-MLLT10 positive acute leukaemia. *Br. J. Haematol.* **157,** 387–390 (2012).
313. Lui, J. C. *et al.* Synthesizing genome-wide association studies and expression microarray reveals novel genes that act in the human growth plate to modulate height. *Hum. Mol. Genet.* **21,** 5193–5201 (2012).
314. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).

315. Khankari, N. K. *et al.* Association between Adult Height and Risk of Colorectal, Lung, and Prostate Cancer: Results from Meta-analyses of Prospective Studies and Mendelian Randomization Analyses. *PLoS Med.* **13,** e1002118 (2016).
316. Bakshi, A. *et al.* Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Sci. Rep.* **6,** 32894 (2016).
317. Taal, H. R. *et al.* Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nat. Genet.* **44,** 532–538 (2012).
318. Ikram, M. A. *et al.* Common variants at 6q22 and 17q21 are associated with intracranial volume. *Nat. Genet.* **44,** 539–544 (2012).
319. Gialluisi, A. *et al.* Genome-wide screening for DNA variants associated with reading and language traits. *Genes. Brain. Behav.* **13,** 686–701 (2014).
320. Liu, F. *et al.* A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS Genet.* **8,** e1002932 (2012).
321. Zhang, Y.-B. *et al.* Genome-wide association study identifies multiple susceptibility loci for craniofacial microsomia. *Nat. Commun.* **7,** 10605 (2016).
322. Shaffer, J. R. *et al.* Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. *PLoS Genet.* **12,** e1006149 (2016).
323. Bachmann, N. *et al.* Mutation screen and association study of EZH2 as a susceptibility gene for aggressive prostate cancer. *Prostate* **65,** 252–259 (2005).
324. Yoon, K.-A., Gil, H. J., Han, J., Park, J. & Lee, J. S. Genetic Polymorphisms in the Polycomb Group Gene EZH2 and the Risk of Lung Cancer. *J. Thorac. Oncol.* **5,** 10–16 (2010).
325. Breyer, J. P. *et al.* Genetic Variants and Prostate Cancer Risk: Candidate Replication and Exploration of Viral Restriction Genes. *Cancer Epidemiol. Biomarkers Prev.* **18,** 2137–2144 (2009).
326. Coulter-Mackie, M. B. & Gagnier, L. Spectrum of mutations in the arylsulfatase A gene in a Canadian DNA collection including two novel frameshift mutations, a new missense mutation (C488R) and an MLD mutation (R84Q) in cis with a pseudodeficiency allele. *Mol. Genet. Metab.* **79,** 91–98 (2003).
327. Yasuda, M. *et al.* Fabry disease: characterization of alpha-galactosidase A double mutations and the D313Y plasma enzyme pseudodeficiency allele. *Hum. Mutat.* **22,** 486–

492 (2003).

328. Tomatsu, S., Montaño, A. M., Dung, V. C., Grubb, J. H. & Sly, W. S. Mutations and polymorphisms in GUSB gene in mucopolysaccharidosis VII (Sly Syndrome). *Hum. Mutat.* **30,** 511–519 (2009).
329. Sarma, K., Margueron, R., Ivanov, A., Pirrotta, V. & Reinberg, D. Ezh2 requires PHF1 to efficiently catalyze H3 lysine 27 trimethylation in vivo. *Mol. Cell. Biol.* **28,** 2718–2731 (2008).
330. Kim, E. *et al.* Phosphorylation of EZH2 Activates STAT3 Signaling via STAT3 Methylation and Promotes Tumorigenicity of Glioblastoma Stem-like Cells. *Cancer Cell* **23,** 839–852 (2013).
331. Sanulli, S. *et al.* Jarid2 Methylation via the PRC2 Complex Regulates H3K27me3 Deposition during Cell Differentiation. *Mol. Cell* **57,** 769–783 (2015).
332. He, A. *et al.* PRC2 directly methylates GATA4 and represses its transcriptional activity. *Genes Dev.* **26,** 37–42 (2012).
333. Lee, J. M. *et al.* EZH2 Generates a Methyl Degron that Is Recognized by the DCAF1/DDB1/CUL4 E3 Ubiquitin Ligase Complex. *Mol. Cell* **48,** 572–586 (2012).
334. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).
335. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42,** 30–35 (2010).
336. Solomon, B. D., Lee, T., Nguyen, A.-D. & Wolfsberg, T. G. A 2.5-year snapshot of Mendelian discovery. *Mol. Genet. Genomic Med.* **4,** 392–394 (2016).
337. Stranneheim, H. & Wedell, A. Exome and genome sequencing: a revolution for the discovery and diagnosis of monogenic disorders. *J. Intern. Med.* **279,** 3–15 (2016).
338. Yang, Y. *et al.* Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *N. Engl. J. Med.* **369,** 1502–1511 (2013).
339. Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312,** 1870–1879 (2014).
340. Xue, Y., Ankala, A., Wilcox, W. R. & Hegde, M. R. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. *Genet. Med.* **17,** 444–451 (2015).

341. Zhu, X. *et al.* Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genet. Med.* **17,** 774–781 (2015).
342. Jamuar, S. S. & Tan, E.-C. Clinical application of next-generation sequencing for Mendelian diseases. *Hum. Genomics* **9,** 10 (2015).
343. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Publ. Gr.* **12,** 745–755 (2011).
344. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
345. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
346. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. *Fly (Austin).* **6,** 80–92 (2012).
347. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
348. Forbes, S. A. *et al.* COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **39,** 945–950 (2011).
349. Fejes, A. P., Khodabakhshi, A. H., Birol, I. & Jones, S. J. M. Human variation database: An open-source database template for genomic discovery. *Bioinformatics* **27,** 1155–1156 (2011).
350. Sewalt, R. G. *et al.* Characterization of interactions between the mammalian polycomb-group proteins Enx1/EZH2 and EED suggests the existence of different mammalian polycomb-group protein complexes. *Mol. Cell. Biol.* **18,** 3586–3595 (1998).
351. Twigg, S. R. F. *et al.* Mutations in multidomain protein MEGF8 identify a carpenter syndrome subtype associated with defective lateralization. *Am. J. Hum. Genet.* **91,** 897–905 (2012).
352. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1,** 40–47 (2000).
353. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).
354. Schumacher, A., Lichtarge, O., Schwartz, S. & Magnuson, T. The Murine Polycomb-

Group Gene eed and Its Human Orthologue: Functional Implications of Evolutionary Conservation. *Genomics* **54,** 79–88 (1998).

355. Han, Z. *et al.* Structural basis of EZH2 recognition by EED. *Structure* **15,** 1306–1315 (2007).
356. Tie, F., Furuyama, T. & Harte, P. J. The Drosophila Polycomb Group proteins ESC and E(Z) bind directly to each other and co-localize at multiple chromosomal sites. *Development* **125,** 3483–3496 (1998).
357. Montgomery, N. D., Yee, D., Montgomery, S. A. & Magnuson, T. Molecular and Functional Mapping of EED Motifs Required for PRC2-Dependent Histone Methylation. *J. Mol. Biol.* **374,** 1145–1157 (2007).
358. Sathe, S. S. & Harte, P. J. The Drosophila extra sex combs protein contains WD motifs essential for its function as a repressor of homeotic genes. *Mech. Dev.* **52,** 77–87 (1995).
359. Ng, J., Li, R., Morgan, K. & Simon, J. Evolutionary conservation and predicted structure of the Drosophila extra sex combs repressor protein. *Mol. Cell. Biol.* **17,** 6663–6672 (1997).
360. Ketel, C. S. *et al.* Subunit contributions to histone methyltransferase activities of fly and worm polycomb group complexes. *Mol. Cell. Biol.* **25,** 6857–6868 (2005).
361. Migliori, V., Mapelli, M. & Guccione, E. On WD40 proteins: propelling our knowledge of transcriptional control? *Epigenetics* **7,** 815–822 (2012).
362. Nangalia, J. *et al.* Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *N. Engl. J. Med.* **369,** 2391–2405 (2013).
363. Kaminsky, E. B. *et al.* An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genet. Med.* **13,** 777–784 (2011).
364. Denisenko, O., Shnyreva, M., Suzuki, H. & Bomsztyk, K. Point mutations in the WD40 domain of Eed block its interaction with Ezh2. *Mol. Cell. Biol.* **18,** 5634–5642 (1998).
365. Ng, J., Hart, C. M., Morgan, K. & Simon, J. A. A Drosophila ESC-E(Z) protein complex is distinct from other polycomb group complexes and contains covalently modified ESC. *Mol. Cell. Biol.* **20,** 3069–3078 (2000).
366. Khan, A. A., Lee, A. J. & Roh, T.-Y. Polycomb group protein-mediated histone modifications during cell differentiation. *Epigenomics* **7,** 75–84 (2015).

367. Cohen, A. S. a *et al.* A novel mutation in EED associated with overgrowth. *J. Hum. Genet.* **60,** 339–342 (2015).
368. Cohen, A. S. A. & Gibson, W. T. EED-associated overgrowth in a second male patient. *J. Hum. Genet.* **61,** 831–834 (2016).
369. Tlemsani, C. *et al.* SETD2 and DNMT3A screen in the Sotos-like syndrome French cohort. *J. Med. Genet.* (2016). doi:10.1136/jmedgenet-2015-103638
370. Eisen, J. A., Sweder, K. S. & Hanawalt, P. C. Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. *Nucleic Acids Res.* **23,** 2715–2723 (1995).
371. Marfella, C. G. A. & Imbalzano, A. N. The Chd family of chromatin remodelers. *Mutat. Res. Mol. Mech. Mutagen.* **618,** 30–40 (2007).
372. Dege, C. & Hagman, J. Mi-2/NuRD chromatin remodeling complexes regulate B and T-lymphocyte development and function. *Immunol. Rev.* **261,** 126–140 (2014).
373. Tong, J. K., Hassig, C. A., Schnitzler, G. R., Kingston, R. E. & Schreiber, S. L. Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex. *Nature* **395,** 917–921 (1998).
374. Denslow, S. A. & Wade, P. A. The human Mi-2/NuRD complex and gene regulation. *Oncogene* **26,** 5433–5438 (2007).
375. Ge, Q., Nilasena, D. S., O'Brien, C. A., Frank, M. B. & Targoff, I. N. Molecular analysis of a major antigenic region of the 240-kD protein of Mi-2 autoantigen. *J. Clin. Invest.* **96,** 1730–1737 (1995).
376. Seelig, H. P. *et al.* The major dermatomyositis-specific Mi-2 autoantigen is a presumed helicase involved in transcriptional activation. *Arthritis Rheum.* **38,** 1389–1399 (1995).
377. Seelig, H. P., Renz, M., Targoff, I. N., Ge, Q. & Frank, M. B. Two forms of the major antigenic protein of the dermatomyositis-specific Mi-2 autoantigen. *Arthritis Rheum.* **39,** 1769–1771 (1996).
378. Jakubaszek, M., Kwiatkowska, B. & Maślińska, M. Polymyositis and dermatomyositis as a risk of developing cancer. *Reumatologia* **53,** 101–105 (2015).
379. da Silva Almeida, A. C. *et al.* The mutational landscape of cutaneous T cell lymphoma and Sézary syndrome. *Nat. Genet.* **47,** 1465–1470 (2015).
380. Kim, M. S., Chung, N. G., Kang, M. R., Yoo, N. J. & Lee, S. H. Genetic and expressional

alterations of CHD genes in gastric and colorectal cancers. *Histopathology* **58,** 660–668 (2011).

381. Kulkarni, S. *et al.* Disruption of chromodomain helicase DNA binding protein 2 (CHD2) causes scoliosis. *Am. J. Med. Genet.* **146A,** 1117–1127 (2008).
382. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380,** 1674–1682 (2012).
383. Carvill, G. L. *et al.* Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nat. Genet.* **45,** 825–830 (2013).
384. Vissers, L. E. L. M. *et al.* Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat. Genet.* **36,** 955–957 (2004).
385. Bernier, R. *et al.* Disruptive CHD8 Mutations Define a Subtype of Autism Early in Development. *Cell* **158,** 263–276 (2014).
386. Dickinson, M. E. *et al.* High-throughput discovery of novel developmental phenotypes. *Nature* **537,** 508–514 (2016).
387. Sparmann, A. *et al.* The chromodomain helicase Chd4 is required for Polycomb-mediated inhibition of astroglial differentiation. *EMBO J.* **32,** 1598–1612 (2013).
388. Weiss, K. *et al.* De Novo Mutations in CHD4, an ATP-Dependent Chromatin Remodeler Gene, Cause an Intellectual Disability Syndrome with Distinctive Dysmorphisms. *Am. J. Hum. Genet.* **99,** 934–941 (2016).
389. Masino, A. J. *et al.* Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology. *BMC Bioinformatics* **15,** 248 (2014).
390. Aronson, S. J. & Rehm, H. L. Building the foundation for genomics in precision medicine. *Nature* **526,** 336–342 (2015).
391. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13,** 565–575 (2012).
392. Haendel, M. A. *et al.* Disease insights through cross-species phenotype comparisons. *Mamm. Genome* **26,** 548–555 (2015).
393. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and

Genomics and the Association for Molecular Pathology. *Genet. Med.* **17,** 405–424 (2015).

394. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).
395. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7,** Unit7.20 (2013).
396. Miosge, L. A. *et al.* Comparison of predicted and actual consequences of missense mutations. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E5189-5198 (2015).
397. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *N. Engl. J. Med.* **372,** 2235–2242 (2015).
398. Javed, A., Agrawal, S. & Ng, P. C. Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat. Methods* **11,** 935–937 (2014).
399. Singleton, M. V. *et al.* Phevor Combines Multiple Biomedical Ontologies for Accurate Identification of Disease-Causing Alleles in Single Individuals and Small Nuclear Families. *Am. J. Hum. Genet.* **94,** 599–610 (2014).
400. Robinson, P. N. *et al.* The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *Am. J. Hum. Genet.* **83,** 610–615 (2008).
401. Buske, O. J. *et al.* PhenomeCentral: A Portal for Phenotypic and Genotypic Matchmaking of Patients with Rare Genetic Diseases. *Hum. Mutat.* **36,** 931–940 (2015).
402. Philippakis, A. A. *et al.* The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum. Mutat.* **36,** 915–921 (2015).
403. Mostafavi, S. & Morris, Q. Combining many interaction networks to predict gene function and analyze gene lists. *Proteomics* **12,** 1687–1696 (2012).
404. Robinson, P. N. *et al.* Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.* **24,** 340–348 (2014).
405. Ayadi, A. *et al.* Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mamm. Genome* **23,** 600–610 (2012).
406. Stelzer, G. *et al.* VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. *BMC Genomics* **17,** 444 (2016).
407. Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical

exome and genome sequencing. *Genet. Med.* **15,** 565–574 (2013).

408. ACMG Board of Directors. ACMG policy statement: updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing. *Genet. Med.* **17,** 68–69 (2015).
409. Bartels, C. F. *et al.* Mutations in the transmembrane natriuretic peptide receptor NPR-B impair skeletal growth and cause acromesomelic dysplasia, type Maroteaux. *Am. J. Hum. Genet.* **75,** 27–34 (2004).
410. Vasques, G. A., Arnhold, I. J. P. & Jorge, A. A. L. Role of the Natriuretic Peptide System in Normal Growth and Growth Disorders. *Horm. Res. Paediatr.* **82,** 222–229 (2014).
411. Olney, R. C. *et al.* Heterozygous Mutations in Natriuretic Peptide Receptor-B ( *NPR2* ) Are Associated with Short Stature. *J. Clin. Endocrinol. Metab.* **91,** 1229–1232 (2006).
412. Hachiya, R. *et al.* Intact Kinase Homology Domain of Natriuretic Peptide Receptor-B Is Essential for Skeletal Development. *J. Clin. Endocrinol. Metab.* **92,** 4009–4014 (2007).
413. Vasques, G. a. *et al.* Heterozygous mutations in natriuretic peptide receptor-B (NPR2) gene as a cause of short stature in patients initially classified as idiopathic short stature. *J. Clin. Endocrinol. Metab.* **98,** 1636–1644 (2013).
414. Amano, N. *et al.* Identification and functional characterization of two novel NPR2 mutations in Japanese patients with short stature. *J. Clin. Endocrinol. Metab.* **99,** E713-8 (2014).
415. Wang, S. R. *et al.* Heterozygous Mutations in Natriuretic Peptide Receptor-B ( *NPR2* ) Gene as a Cause of Short Stature. *Hum. Mutat.* **36,** 474–481 (2015).
416. Miura, K. *et al.* An overgrowth disorder associated with excessive production of cGMP due to a gain-of-function mutation of the natriuretic peptide receptor 2 gene. *PLoS One* **7,** e42180 (2012).
417. Hannema, S. E. *et al.* An activating mutation in the kinase homology domain of the natriuretic peptide receptor-2 causes extremely tall stature without skeletal deformities. *J. Clin. Endocrinol. Metab.* **98,** 1988–1998 (2013).
418. Miura, K. *et al.* Overgrowth syndrome associated with a gain-of-function mutation of the natriuretic peptide receptor 2 (NPR2) gene. *Am. J. Med. Genet.* **164A,** 156–163 (2014).
419. Franco, L. M. *et al.* A syndrome of short stature, microcephaly and speech delay is associated with duplications reciprocal to the common Sotos syndrome deletion. *Eur. J.*

*Hum. Genet.* **18,** 258–261 (2010).

420. Rosenfeld, J. A. *et al.* Further evidence of contrasting phenotypes caused by reciprocal deletions and duplications: Duplication of NSD1 causes growth retardation and microcephaly. *Mol. Syndromol.* **3,** 247–254 (2012).
421. Dikow, N. *et al.* The phenotypic spectrum of duplication 5q35.2-q35.3 encompassing NSD1: Is it really a reversed sotos syndrome? *Am. J. Med. Genet.* **161A,** 2158–2166 (2013).
422. Žilina, O. *et al.* Patient with Dup(5)(q35.2-q35.3) reciprocal to the common Sotos syndrome deletion and review of the literature. *Eur. J. Med. Genet.* **56,** 202–206 (2013).
423. Novara, F. *et al.* Defining the phenotype associated with microduplication reciprocal to Sotos syndrome microdeletion. *Am. J. Med. Genet.* **164A,** 2084–2090 (2014).
424. Fernandez, B. A. *et al.* Adult siblings with homozygous G6PC3 mutations expand our understanding of the severe congenital neutropenia type 4 (SCN4) phenotype. *BMC Med. Genet.* **13,** 111 (2012).
425. Adams, D. R. *et al.* Three rare diseases in one Sib pair: RAI1, PCK1, GRIN2B mutations associated with Smith–Magenis Syndrome, cytosolic PEPCK deficiency and NMDA receptor glutamate insensitivity. *Mol. Genet. Metab.* **113,** 161–170 (2014).
426. Posey, J. E. *et al.* Molecular diagnostic experience of whole-exome sequencing in adult patients. *Genet. Med.* **18,** 678–685 (2016).
427. Shashi, V. *et al.* The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genet. Med.* **16,** 176–182 (2014).
428. Thevenon J *et al.* Diagnostic odyssey in severe neurodevelopmental disorders: Towards clinical whole-exome sequencing as a first-line diagnostic test. *Clin. Genet.* **89,** 700–707 (2016).
429. Cao, Q. *et al.* The central role of EED in the orchestration of polycomb group complexes. *Nat. Commun.* **5,** 3127 (2014).
430. Margueron, R. *et al.* Ezh1 and Ezh2 Maintain Repressive Chromatin through Different Mechanisms. *Mol. Cell* **32,** 503–518 (2008).
431. Shen, X. *et al.* EZH1 Mediates Methylation on Histone H3 Lysine 27 and Complements EZH2 in Maintaining Stem Cell Identity and Executing Pluripotency. *Mol. Cell* **32,** 491–

502 (2008).

432. van Asperen, C. J., Overweg-Plandsoen, W. C., Cnossen, M. H., van Tijn, D. A. & Hennekam, R. C. Familial neurofibromatosis type 1 associated with an overgrowth syndrome resembling Weaver syndrome. *J. Med. Genet.* **35,** 323–327 (1998).
433. Pasmant, E. *et al.* NF1 microdeletions in neurofibromatosis type 1: from genotype to phenotype. *Hum. Mutat.* **31,** E1506–E1518 (2010).
434. Ning, X. *et al.* Growth in neurofibromatosis 1 microdeletion patients. *Clin. Genet.* **89,** 351–354 (2016).
435. Cohen, A. S. A., Wilson, S. L., Trinh, J. & Ye, X. C. Detecting somatic mosaicism: Considerations and clinical implications. *Clin. Genet.* **87,** 554–562 (2015).
436. Saugier-Veber, P. *et al.* Heterogeneity of NSD1 alterations in 116 patients with Sotos syndrome. *Hum. Mutat.* **28,** 1098–1107 (2007).
437. Castronovo, C. *et al.* A novel mosaic NSD1 intragenic deletion in a patient with an atypical phenotype. *Am. J. Med. Genet.* **161A,** 611–618 (2013).
438. Alders, M. *et al.* Methylation analysis in tongue tissue of BWS patients identifies the (EPI)genetic cause in 3 patients with normal methylation levels in blood. *Eur. J. Med. Genet.* **57,** 293–297 (2014).
439. Chen, H. *et al.* Wedelolactone disrupts the interaction of EZH2-EED complex and inhibits PRC2-dependent cancer. *Oncotarget* **6,** 13049–13059 (2015).
440. Yan, W., Herman, J. G. & Guo, M. Epigenome-based personalized medicine in human cancer. *Epigenomics* **8,** 119–133 (2015).
441. Schapira, M. & Arrowsmith, C. H. Methyltransferase inhibitors for modulation of the epigenome and beyond. *Curr. Opin. Chem. Biol.* **33,** 81–87 (2016).
442. Ronan, J. L., Wu, W. & Crabtree, G. R. From neural development to cognition: unexpected roles for chromatin. *Nat. Rev. Genet.* **14,** 347–359 (2013).
443. Pennisi, E. The CRISPR craze. *Science* **341,** 833–836 (2013).
444. Kato, T. & Takada, S. In vivo and in vitro disease modeling with CRISPR/Cas9. *Brief. Funct. Genomics* (2016). doi:10.1093/bfgp/elw031
445. Wassef, M., Michaud, A. & Margueron, R. Association between EZH2 expression, silencing of tumor suppressors and disease outcome in solid tumors. *Cell Cycle* **15,** 2256–2262 (2016).

446. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31,** 822–826 (2013).
447. Tsai, S. Q. & Joung, J. K. Defining and improving the genome-wide specificities of CRISPR-Cas9 nucleases. *Nat. Rev. Genet.* **17,** 300–312 (2016).
448. Choufani, S. *et al.* NSD1 mutations generate a genome-wide DNA methylation signature. *Nat. Commun.* **6,** 10207 (2015).
449. Rose, N. R. & Klose, R. J. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim. Biophys. Acta* **1839,** 1362–1372 (2014).
450. Visser, R., Hasegawa, T., Niikawa, N. & Matsumoto, N. Analysis of the NSD1 promoter region in patients with a Sotos syndrome phenotype. *J. Hum. Genet.* **51,** 15–20 (2006).
451. Shankar, G. M. *et al.* Sporadic hemangioblastomas are characterized by cryptic VHL inactivation. *Acta Neuropathol. Commun.* **2,** 167 (2014).
452. Papaemmanuil, E. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122,** 3616–27; quiz 3699 (2013).
453. Shain, A. H. *et al.* Exome sequencing of desmoplastic melanoma identifies recurrent NFKBIE promoter mutations and diverse activating mutations in the MAPK pathway. *Nat. Genet.* **47,** 1194–1199 (2015).
454. De Keersmaecker, K. *et al.* Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nat. Genet.* **45,** 186–90 (2013).
455. Lasho, T. L. *et al.* Novel recurrent mutations in ethanolamine kinase 1 (ETNK1) gene in systemic mastocytosis with eosinophilia and chronic myelomonocytic leukemia. *Blood Cancer J.* **5,** e275 (2015).
456. Xu, L. *et al.* Genomic landscape of CD34+ hematopoietic cells in myelodysplastic syndrome and gene mutation profiles as prognostic markers. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 8589–8594 (2014).
457. Douglas, J. *et al.* Evaluation of NSD2 and NSD3 in overgrowth syndromes. *Eur. J. Hum. Genet.* **13,** 150–153 (2005).
458. Qian, C. & Zhou, M. M. SET domain protein lysine methyltransferases: Structure, specificity and catalysis. *Cell. Mol. Life Sci.* **63,** 2755–2763 (2006).
459. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond.

*Nat. Rev. Genet.* **13,** 484–92 (2012).

460. Woo, C. J., Kharchenko, P. V., Daheron, L., Park, P. J. & Kingston, R. E. A Region of the Human HOXD Cluster that Confers Polycomb-Group Responsiveness. *Cell* **140,** 99–110 (2010).
461. Kim, J. & Kim, H. Recruitment and biological consequences of histone modification of H3K27me3 and H3K9me3. *ILAR J.* **53,** 232–239 (2012).
462. Iwabuchi, K. *et al.* Stimulation of p53-mediated transcriptional activation by the p53-binding proteins, 53BP1 and 53BP2. *J. Biol. Chem.* **273,** 26061–26068 (1998).
463. Dimitrova, N., Chen, Y.-C. M., Spector, D. L. & de Lange, T. 53BP1 promotes non-homologous end joining of telomeres by increasing chromatin mobility. *Nature* **456,** 524–528 (2008).
464. Fradet-Turcotte, A. *et al.* 53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark. *Nature* **499,** 50–54 (2013).
465. Wilson, M. D. *et al.* The structural basis of modified nucleosome recognition by 53BP1. *Nature* **536,** 100–103 (2016).
466. Grozeva, D. *et al.* De novo loss-of-function mutations in SETD5, encoding a methyltransferase in a 3p25 microdeletion syndrome critical region, cause intellectual disability. *Am. J. Hum. Genet.* **94,** 618–624 (2014).
467. Tchasovnikarova, I. A. *et al.* GENE SILENCING. Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells. *Science* **348,** 1481–1485 (2015).
468. Fei, Q. *et al.* SETDB1 modulates PRC2 activity at developmental genes independently of H3K9 trimethylation in mouse ES cells. *Genome Res.* **25,** 1325–1335 (2015).
469. Jiang, H. *et al.* Regulation of transcription by the MLL2 complex and MLL complex-associated AKAP95. *Nat. Struct. Mol. Biol.* **20,** 1156–1163 (2013).
470. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42,** 790–793 (2010).
471. Fahrner, J. A. & Bjornsson, H. T. Mendelian Disorders of the Epigenetic Machinery: Tipping the Balance of Chromatin States. *Annu. Rev. Genomics Hum. Genet.* **15,** 269–293 (2014).

# Appendices

## Appendix A: Distribution of somatic *EZH2* alterations across tissues, according to the COSMIC database (July 2016)

| Tissue | Point mutations | Copy number variants | Gene expression changes |
|---|---|---|---|
| Adrenal gland | 280 | - | 79 |
| Autonomic ganglia | 748 | - | - |
| Biliary tract | 329 | - | - |
| Bone | 535 | - | - |
| Breast | 1,540 | 997 | 1,104 |
| Central nervous system | 2,221 | 813 | 695 |
| Cervix | 322 | 176 | 307 |
| Endometrium | 640 | 426 | 602 |
| Eye | 39 | - | - |
| Fallopian tube | 2 | - | - |
| Genital tract | 29 | - | - |
| Haematopoietic and lymphoid | 10,810 | 731 | 221 |
| Kidney | 1,679 | 428 | 600 |
| Large intestine | 1,578 | 700 | 607 |
| Liver | 1,628 | 657 | 373 |
| Lung | 1,920 | 1,114 | 1,018 |
| Meninges | 65 | - | - |
| Oesophagus | 1,109 | - | 125 |
| Ovary | 843 | 721 | 266 |
| Pancreas | 1,594 | 706 | 179 |
| Parathyroid | 218 | - | - |
| Peritoneum | 10 | - | - |
| Pituitary | 15 | - | - |
| Placenta | 2 | - | - |
| Pleura (lung membrane) | 77 | - | - |
| Prostate | 1,303 | 303 | 498 |
| Salivary gland | 147 | - | - |
| Skin | 1,168 | 486 | 472 |
| Small intestine | 43 | - | - |
| Soft tissue | 671 | 143 | 263 |
| Stomach | 713 | - | 285 |
| Testis | 20 | - | - |
| Thymus | 52 | - | - |
| Thyroid | 630 | 366 | 513 |
| Upper aerodigestive tract | 1,152 | 468 | 522 |
| Urinary tract | 667 | - | 408 |
| Vulva | 3 | - | - |
| *Not specific* | 54 | 11 | - |
| **TOTAL** | **34,856** | **9,246** | **9,137** |

## Appendix B: Variation reported in *EZH2*, according to the dbSNP database (June 2015)

| Genomic coordinates (chr7, GRCh37.p13) | Predicted protein change | exon | dbSNP reference | PROVEAN prediction | SIFT prediction | Minor allele frequency/number of cases identified | Source of data & Notes |
|---|---|---|---|---|---|---|---|
| 7,148544361,C,G | K10N | 2 | rs764649740 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148544360,C,A | G11* | 2 | rs756498768 | STOP | STOP | *single case* | Decode in Iceland – genomic, not validated |
| 7,148544347,C,T | W15* | 2 | rs760133156 | STOP | STOP | *single case* | ExAc – genomic, not validated |
| 7,148544342,T,G | K17Q | 2 | rs776984937 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148544339,G,A | R18C | 2 | rs771352080 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148543690,T,C | S40G | 3 | rs754403133 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148543687,T,C | M41V | 3 | rs766928732 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148543677,G,T | S44Y | 3 | rs761247972 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148543660,A,C | L50V | 3 | rs772530358 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148543659,A,G | L50S | 3 | rs775407864 | neutral | tolerated | *3 cases* | ExAc, UK10KTwins & UK10K – genomic, not validated |
| 7,148543653,C,T | R52K | 3 | rs752284693 | neutral | tolerated | *2 cases* | ExAc & Decode– genomic, not validated |
| 7,148543650,G,C | T53R | 3 | rs768812143 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148543643,G,C | I55M | 3 | rs199645805 | neutral | tolerated | *6 cases: 0.001-0.002* | various |
| 7,148543630,A,G | W60R | 3 | rs746946161 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148543576,G,A | R78C | 3 | rs141583753 | neutral | damaging | *2 cases* | ExAc & NHLBI project – genomic, not validated |
| 7,148543567,T,C | R81G | 3 | rs765980265 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148529820,T,A | D90V | 4 | rs765323739 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148529815,G,A | P92S | 4 | rs759576587 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148529811,G,C | T93R | 4 | rs773141417 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148529791,T,A | T100S | 4 | rs748108870 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148526919,G,A | H129Y | 5 | rs189454324 | deleterious | damaging | *4 cases: 0.0002/1* | 1000G (3 phases) & ExAc – genomic, not validated - UNLIKELY TO BE PATHOGENIC |
| 7,148526910,G,A | P132S | 5 | rs193921148 | deleterious | damaging | *ours + 1* | added by D. Bulman, Ottawa; Gibson *et al.*[56]; there is also a common SYNONYMOUS variant at this site that is non-pathogenic (see rs61732845 at G=0.0016/8) |
| 7,148526892,C,T | V138I | 5 | rs766447146 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148526845_148526847delATA | Y153del | 5 | rs193921146 | - | - | *ours + 1* | added by D. Bulman, Ottawa; Gibson *et al.*[56] |
| 7,148526840,C,T | G155E | 5 | rs375168091 | deleterious | tolerated | *2 cases* | ExAc & NHLBI – genomic, not validated |
| 7,148525961,T,A | I166L | 6 | rs759313335 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148525912,T,C | N182S | 6 | rs746465165 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148525902,GTCA,G | D183del | 6 | rs766699965 | neutral | N/A | *3 cases* | ExAc, UK10KTwins & UK10K – genomic, not validated |
| 7,148525893,ATCATCATCG,A | D184_187del (DDD) | 6 | rs761287299 | deleterious | N/A | *single case* | ExAc – genomic, not validated |
| **7,148525904,C,G** | **D185H** | **6** | **rs2302427** | **neutral** | **damaging** | ***0.0799/400*** | various: **MOST COMMON VARIANT**- UNLIKELY TO BE PATHOGENIC; there is also a SYNONYMOUS variant at this site: rs555589547 |
| 7,148525888,CCAT,C | D188del | 6 | rs587778303 | neutral | N/A | *2 cases* | ExAc – genomic, not validated; also reported in ClinVar |

| Genomic coordinates (chr7, GRCh37.p13) | Predicted protein change | exon | dbSNP reference | PROVEAN prediction | SIFT prediction | Minor allele frequency/number of cases identified | Source of data & Notes |
|---|---|---|---|---|---|---|---|
| | | | | | | | (as "untested allele") |
| 7,148525888,CCAT,CCATCAT | D188ins (+D) | 6 | rs751123994 | neutral | N/A | *single case* | ExAc – genomic, not validated |
| 7,148525883,C,T | D192N | 6 | rs778968366 | neutral | tolerated | *3 cases* | ExAc, UK10KTwins & UK10K– genomic, not validated |
| 7,148525867,T,C | E197G | 6 | rs749498698 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148525864,T,G | E198A | 6 | rs780251816 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148525853,C,T | D202N | 6 | rs192731117 | neutral | tolerated | *4 cases: 0.0004/2* | 1000G (3 phases) & ExAc – genomic, not validated - UNLIKELY TO BE PATHOGENIC |
| 7,148525851,A,C | D202E | 6 | rs767489605 | neutral | tolerated | *2 cases* | ExAc – genomic, not validated |
| 7,148525849,A,C | L203R | 6 | rs758449513 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148525837,C,T | R207Q | 6 | rs765147666 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148525834,T,C | D208G | 6 | rs61753264 | neutral | tolerated | *3 cases + cancer* | ExAc, NHLBI, etc – genomic, not validated |
| 7,148524357,A,T | D209E | 7 | rs770030757 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148524347,G,A | R213C | 7 | rs112029831 | neutral | tolerated | *3 cases* | ExAc, NHBLI & Bushman – genomic, not validated - UNLIKELY TO BE PATHOGENIC |
| 7,148524346,C,T | R213H | 7 | rs377467108 | neutral | tolerated | *2 cases* | ExAc & NHLBI – genomic, not validated |
| 7,148524338,G,A | R216W | 7 | rs771139896 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148524337,C,T | R216Q | 7 | rs747028969 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148524325,G,A | S220F | 7 | rs537373788 | deleterious | tolerated | *single case* | Genome of the Netherlands – genomic, not validated |
| 7,148524323,C,G | D221H | 7 | rs374699518 | deleterious | tolerated | *single case* | NHBLI – genomic, not validated; there is also a SYNONYMOUS variant at this site: rs755014667 |
| 7,148524296,T,C | M230V | 7 | rs779951996 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148524277,G,A | T236I | 7 | rs755823072 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148523705,G,C | Q250E | 8 | rs200520401 | neutral | tolerated | *0.002* | ClinSeq case & CSAgilent "population" |
| 7,148523689,G,A | A255V | 8 | rs372285596 | neutral | tolerated | *2 cases* | ExAc & NHLBI – genomic, not validated |
| 7,148523687,G,C | L256V | 8 | rs568618347 | neutral | tolerated | *single case* | Genome of the Netherlands – genomic, not validated |
| 7,148523666,T,C | N263D | 8 | rs780959381 | deleterious | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148523663,T,C | I264V | 8 | rs756942104 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148523636,G,T | Q273K | 8 | rs763682209 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148523617,T,G | H279P | 8 | rs757807865 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148523605,G,A | T283M | 8 | rs587778304 | deleterious | damaging | *single case* | ClinVar ("untested allele") – genomic, not validated |
| 7,148516755,T,G | Y311S | 9 | rs766875832 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148516735,T,C | T318A | 9 | rs773492931 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148516722,T,C | N322S | 9 | rs151023145 | neutral | tolerated | *0.0002/1* | various including ClinVar ("untested allele"); there is also a SYNONYMOUS variant at this site: rs768699214 |
| 7,148516690,A,C | L333V | 9 | rs746749718 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148515187,G,A | A341V | 10 | rs747782211 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148515175,G,A | A345V | 10 | rs748860527 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148515154,G,A | P352L | 10 | rs200964386 | neutral | damaging | *3 cases* | 1000G (2 phases) & ExAc – genomic, not validated |
| 7,148515148,C,G | R354P | 10 | rs775942317 | neutral | tolerated | *2 cases* | UK10KTwins & UK10K – genomic, not validated |
| 7,148515124,C,T | R362Q | 10 | rs781468426 | neutral | tolerated | *2 cases* | ExAc & Finnish study – genomic, not validated |

| Genomic coordinates (chr7, GRCh37.p13) | Predicted protein change | exon | dbSNP reference | PROVEAN prediction | SIFT prediction | Minor allele frequency/number of cases identified | Source of data & Notes |
|---|---|---|---|---|---|---|---|
| 7,148515122,G,A | L363F | 10 | rs781431240 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148515118,G,A | P364L | 10 | rs757601923 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148515074,C,G | E379Q | 10 | rs144316514 | neutral | tolerated | *single case* | NHLBI- Genomic-not validated |
| 7,148515061,G,A | T383I | 10 | rs553185801 | neutral | tolerated | *2 cases: 0.0002/1* | ExAc & 1000G (3rd phase) – genomic, not validated |
| 7,148515028,C,G | G394A | 10 | rs587778305 | neutral | tolerated | *single case* | ClinVar ("untested allele") – genomic, not validated |
| 7,148515005,CTTCTTT,C | (KE)400del | 10 | rs756400659 | neutral | N/A | *single case* | ExAc – genomic, not validated |
| 7,148515008,CTTT,C | K400del | 10 | rs780357774 | neutral | N/A | *single case* | ExAc – genomic, not validated |
| 7,148515010,T,G | K400T | 10 | rs774270705 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148514997,CTCT,C | E402del | 10 | rs775039041 | neutral | N/A | *3 cases* | ExAc, UK10KTwins & UK10K – genomic, not validated |
| 7,148514993,T,G | K406Q | 10 | rs748915411 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148514989,T,C | D407G | 10 | rs779757594 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148514986,T,C | E408G | 10 | rs769298548 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148514466,G,C | Q420E | 11 | rs775295296 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148514442,T,G | N428H | 11 | rs776246781 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148514438,A,G | I429T | 11 | rs770327434 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148514393,A,G | M444T | 11 | rs370444695 | deleterious | damaging | *2 cases* | ExAc & NHLBI – genomic, not validated |
| 7,148514354,C,T | C457Y | 11 | rs748458685 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148514349,T,G | I459L | 11 | rs779136188 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148513845,G,T | S479Y | 12 | rs765768601 | neutral | damaging | *2 cases* | UK10K & UK10KTwins – genomic, not validated |
| 7,148513840,T,G | I481L | 12 | rs566622851 | neutral | tolerated | *single case* | 1000G phase 3 – genomic, not validated |
| 7,148513822,C,T | A487T | 12 | rs201135441 | neutral | tolerated | *0.0012/6* | various including ClinVar ("untested allele") |
| 7,148513807,T,A | T492S | 12 | rs770006533 | deleterious | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148513780,G,A | H501Y | 12 | rs746001943 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148512129,C,T | G517S | 14 | rs776925814 | neutral | tolerated | *single case* | ExAc – genomic, not validated; there is also a SYNONYMOUS variant at this site: rs749179228 |
| 7,148512123,A,G | S519P | 14 | rs747009766 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148512079,C,A | Q533H | 14 | rs768512235 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148512078,G,A | P534S | 14 | rs749239554 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148512025,A,T | F551L | 14 | rs745554458 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148511164,G,C | L580V | 15 | rs754291699 | deleterious | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148511137,G,A | L589F | 15 | rs767528515 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148511131,G,C | L591V | 15 | rs761834990 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148511116,C,T | A596T | 15 | rs139878257 | neutral | tolerated | *0.001/1* | various |
| 7,148511099,A,C | S601R | 15 | rs147328633 | neutral | tolerated | *single case* | NHLBI – genomic, not validated |
| 7,148511092,C,A | V604L | 15 | rs587778302 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148511092,C,T | V604M | 15 | rs587778302 | neutral | damaging | *single case* | ClinVar ("untested allele") – genomic, not validated |
| 7,148511064,C,T | R613Q | 15 | rs561605379 | deleterious | damaging | *2 cases: 0.0002/1* | ExAc & 1000G (3rd phase) – genomic, not validated |
| 7,148511059,A,G | S615P | 15 | rs112034331 | neutral | damaging | *single case* | Bushman "normal population" |
| 7,148511058,G,A | S615F | 15 | rs587778301 | neutral | damaging | *single case* | ClinVar ("untested allele") – genomic, not validated |
| 7,148508788,C,T | V626M | 16 | rs587783625 | neutral | damaging | *single case* | ClinVar "pathogenic" from U.Chicago |

| Genomic coordinates (chr7, GRCh37.p13) | Predicted protein change | exon | dbSNP reference | PROVEAN prediction | SIFT prediction | Minor allele frequency/number of cases identified | Source of data & Notes |
|---|---|---|---|---|---|---|---|
| 7,148508773,T,C | I631V | 16 | rs781407066 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148508767,T,A | I633F | 16 | rs757534855 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148508765,G,C | I633M | 16 | rs751723382 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148508743,C,G | E641Q | 16 | rs753739962 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148508728,A,G | Y646H | 16 | rs267601395 | deleterious | damaging | - | Cancer hotspot |
| 7,148508727,T,A | Y646F | 16 | rs267601394 | deleterious | damaging | - | Cancer hotspot |
| 7,148508719,C,T | E649K | 16 | rs766387427 | deleterious | damaging | *single case* | ExAc – genomic, not validated; there is also a SYNONYMOUS variant at this site: rs760495918 |
| 7,148507463,T,C | D664G | 17 | rs761656628 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148506468,C,T | A682T | 18 | rs397515547 | deleterious | damaging | *single case* | ClinVar "pathogenic": Tatton-Brown *et al*.[58] |
| 7,148506462,G,A | R684C | 18 | rs587783626 | deleterious | damaging | *single case* | ClinVar "pathogenic": Tatton-Brown *et al*.[58]+proband 5 |
| 7,148506432,G,A | H694Y | 18 | rs193921147 | deleterious | damaging | *single case* | ClinVar "pathogenic": our patient from Gibson *et al*.[56] |
| 7,148506426,C,G | V696L | 18 | rs781275057 | neutral | damaging | *single case* | ExAc – genomic, not validated |
| 7,148506426,C,T | V696I | 18 | rs781275057 | neutral | tolerated | *3 cases* | ExAc, UK10KTwins & UK10K – genomic, not validated |
| 7,148506402,C,T | V704I | 18 | rs771467281 | neutral | tolerated | *single case* | ExAc – genomic, not validated |
| 7,148506225,G,GTTA | H711ins(N) | 19 | rs752311892 | deleterious | N/A | *single case* | ExAc – genomic, not validated |
| 7,148506185,C,T | E725K | 19 | rs747933788 | deleterious | damaging | *single case* | ExAc – genomic, not validated |
| 7,148504761,C,T | E745K | 20 | rs397515548 | deleterious | damaging | *single case* | ClinVar "pathogenic": Tatton-Brown *et al*.[58] +proband 10 |
| 7,148504758,T,C | R746G | 20 | rs587783627 | deleterious | damaging | *single case* | ClinVar "likely pathogenic" from U.Chicago |
| 7,148504744,G,C | I750M | 20 | rs779629814 | neutral | tolerated | *2 cases* | ExAc & Spain; there is also a SYNONYMOUS variant at this site reported once in ExAc |

Bright yellow = most common variants; light yellow = variants reported in 2-3 cases; white = variants reported only once; dark green = confirmed pathogenic variants; light green = likely pathogenic variants; purple = variants affecting the cancer "hotspot" amino acid Tyrosine 646.

## Appendix C: PCR and sequencing primers for Sanger sequencing

### C.1 Primers for *EZH2* amplification and sequencing

| Exon | Forward primer (antisense to ORF)* | Reverse primer (sense to ORF)* | Product size (bp) |
|---|---|---|---|
| 2 | CAGATCAAGAACCTAAGCTTCCA | TAGTTTGCTGCGGATTAAAACA | 358 |
| 3 | GACACCCTGAGGTCAATGATTT | TTTTTAACCTGCTTTACAGGTGTT | 314 |
| 4 | TCTTGATTCACCTTGACAATAAAA | ATTTGGGTAGGCAGCATCTCT | 300 |
| 5 | TGCCCTATATGCTTCATAAACAA | TGGGTAAAGACATGTACACATGAAA | 300 |
| 6 | CCTGGCCATAATATGTTAATTTTG | ACTAGGCTATGCCTGTTTTGTCC | 397 |
| 7 | GCTCATCCGCTACATTGATT | AGAAAATCAGCTTTGTTATAGAGACAT | 300 |
| 8 | AAATGATAGCACTCTCCAAGCTG | GCCATTCCTTTATGTTTTAGGC | 371 |
| 9 | AGCATGGGTGCAGACAACAT | TCCATTAATTGACTTTTCCAGTG | 282 |
| 10 | TCTGGTCTTTATACTGAAACTAACCAA | GATTATTTGTGATAAATGGATAATGTG | 489 |
| 11 | TTTTTAGGAGATGAATAGGAGCTT | TGTCCTCATCTTTTCGCTTTT | 390 |
| 12 | CCAACAACAGCCCTTAGGAA | CCCAGCATCTAGCAGTGTCA | 294 |
| 13 | AACCCAAGCTCTAATCCAGTTA | TCTTGGCTTTAACGCATTCC | 250 |
| 14 | AGGGAGTGCTCCCATGTTCT | GCCAGCTACACTCCACAGGT | 331 |
| 15 | TTTGCCCCAGCTAAATCATC | GTACAGCCCTTGCCACGTAT | 351 |
| 16 | TCCAATCAAACCCACAGACTT | TGAGGATTTACAGTGATAGCTTTTG | 299 |
| 17 | CCTCTACCCTCGTTTCTGAACA | CTTGGCTGTAGTGACCCTTTTT | 300 |
| 18 | GGGGGTTAACTGACTTGTTCAC | AGGCAAACCCTGAAGAACTGTA | 297 |
| 19 | GGCAAAGTGACCCATCAAAA | TGGACTTGAATACTTCTGGGATA | 300 |
| 20 | CACTTTGCAGCTGGTGAGAA | TGCACCCACTATCTTCAGCA | 342 |
| 20 seq (FFi/RRi) | GGAGGTAGCAGATGTCAAG | AGCACATGTTGGATGGGT | (129) |

* All sequences are provided in the 5' to 3' orientation. Please note that the *EZH2* gene runs in the opposite orientation from the Reference genome (on the minus strand).
ORF = Open reading frame; bp = Base pairs; NA = Non applicable.
Black = primers from Gibson *et al.*[56]; blue = newly designed primers.

## C.2 Primers for *NSD1* amplification and sequencing

| Exon | Forward primer (sense to ORF)* | Reverse primer (antisense to ORF)* | Product size (bp) |
|---|---|---|---|
| 1 | GCTGCTGCCTCCATTTTGTTT | CGGCCTCCATCTTAGGTTACA | 263 |
| 2.1 | AGAGTCGAGTCAGATGGCCTA | GATCCATCAGCAGACCCATT | 355 |
| 2.2 | GTGGAACATCCCAAAATGCT | TCTGTGACTGGCTGTTCTGG | 367 |
| 2.3 | TGGCTTTCTGCACTTTGAGA | GAAGGGCTGCTTTTTCATTG | 317 |
| 2.4 | GCCATTCTTGCCATTAGCTC | TTTCCCTTTAAGTGGCCTGT | 323 |
| 3 | TGCTTTTTCAGAAGGCTAATAGG | TCATTCACAAAATGTTCCAAGG | 332 |
| 4 | GGTTGTCTAGTTCAGTGGGCA | TCCAATCTGGGAAACAGAGC | 417 |
| 5.1 | TCTGATTTCATCTCCCTTTTCC | CTGTGAGGCTATTTGCTATCC | 515 |
| 5.2 | ATGCCATTTGAAGACTGCAC | TCCACAGGAAGAAAACAGAAAA | 360 |
| 5.3 | TCCAGAGAACCTTGGCCTAAAC | TCCAGGCTCTGCACTCTTAG | 580 |
| 5.4 | GGAAAAGCGAAGTGATTCCA | TCTGACTGGGGTTTGTGAAC | 348 |
| 5.5 | GAAGCCTCTCATTAGTAACTC | ATGGCTTTGATGTTCCAGAG | 554 |
| 5.6 | CAAASAGCCCAAGTTCCGAAGT | AGACAAMTYGCCAGATAATGC | 631 |
| 5.7 | ATCCGAGTTGAAGGAACTCTC | CACACTTGGAAGCTGATTCAG | 604 |
| 5.8 | CCTGTAGGAGTCTCTAAGGT | CACCSCTTTTTRGGCACCAC | 635 |
| 5.9 | CTTCATCCAAATTGCGAGATGC | CAAGTATGCTTGCTGAAGGAG | 567 |
| 5.10 | ACCTCGTAAGCGCATGAACAG | CTTCACTTTACCATTACAACAGACC | 387 |
| 6 | ATGTGGTTTCCCATCTGGTT | AGTACTGTGCTAGAAGCTGAGA | 374 |
| 7 | TGTCTTCAAGGTTCATCCAC | ATTTCCCAGGACAAAAGGGGG | 532 |
| 8 | TACCATCCTGCCTCTTCCCAT | TGACTGCTGACACACACACTA | 379 |
| 9 | TGGCAGCTGACAATTCAGAC | CTCACTGGTCGGGCTTACAC | 268 |
| 10 | CCCGTTTTCCTAATCCACAA | CCTCTGGCGTGAAAAGTAGC | 310 |
| 11 | ACAGCCTCAGAGCAGTTAGT | TATATTCCTTCATGGGCCTTAAG | 590 |
| 12 | CTACAACTACGGGCCCTTGC | TGGCATCAGCTCATCTTTGCT | 517 |
| 13 | TGGGTTCAGACGATGTCAAA | TGACATGGTGGATTGATTGCAT | 743 |
| 14 | TCCATCATCTTAGTGGTCATTCC | CATTAATTCGGGACATCATTTC | 471 |
| 15 | GAGCTAGAATTCCAACCTGAAC | GCGCCTGCCTTGTGAGATTATT | 651 |
| 16 | TGTGGACAGACAGACATTGCT | TTTGCAGCCAGATAATGCCC | 443 |
| 17 | TACCCTTTGGACTACATTACCTGT | GAACTATGCTTGTGCTTCCGTT | 399 |
| 18 | ATGGGAAATGTGGCTGCAACT | ACCTCCTACACAGTGACCATGA | 598 |
| 19 | CTGGGGCRGGTAGAGATTGG | TGGGAACTGCGTTACATGCC | 524 |
| 20 | TGTGCATCTGGGTTGAAACT | GTGGTGATGGTTGCACAAAA | 669 |
| 21 | TCTCTTGGGAGTTGGTATCC | TTTCTTCTTAAAATGTAGACTGCCC | 283 |
| 22.1 | ACCAGCCTTATGGATCAGCA | TGCGTGGTAAACTTTTGGGC | 420 |
| 22.2 | GTGTTCACAGAATGCTGACTG | GAAACACAGCAAGTGCACCG | 458 |
| 23.1 | TGGTGAGTGGCATAAGCTCT | GGAGGCACATACTCACGGAT | 349 |
| 23.2 | CTGTGTTTTCAGGGAAATGGGA | GTGGTTGGGACCTGACTGAG | 597 |
| 23.3 | AGAGCAATCAACAGGAATGGC | CTGATAGTACTTTCTCAGGAGG | 569 |
| 23.4 | GACAAACCCTCTCCAGTGAC | TCATCAGCCTGTGGTGTGAC | 452 |
| 23.5-6 | CAGACTTCAGACAGGCCTACT | CAGGGACTTTGCTCTGTGGT | 674 |
| 23.7 | TTTCTCAGCCTCCTGCCAAG | TCACACAACCATAAGCCCCC | 666 |

* All sequences are provided in the 5' to 3' orientation. ORF = Open reading frame; bp = Base pairs.
Please note that exons 2, 5, 22 and 23 were too large to amplify in a single reaction and thus were subdivided into smaller fractions (with sufficient overlap to analyze the entire coding sequence of each exon).
Black = primers from Douglas *et al.*[283]; red = primers from Rio *et al.*[130] ; blue = newly designed primers.

## C.3 Primers for *EED* amplification and sequencing

| Exon | Forward primer (sense to ORF)* | Reverse primer (antisense to ORF)* | Product size (bp) |
|---|---|---|---|
| 1 | TGGCTGTAACTCATTGGAGTCT | CCGCGTTTTCCTGAGTGACA | 1250 |
| 1a seq | GAGGAGGCGGGTTTCGA | GCGCGGCTTCCTGGC | (545) |
| 1b seq | ATCGTGTAGACTGCCGGGA | *same as exon 1 PCR reverse primer* | (532) |
| 2 | ACCTTCTAACCTGTAGCTGGA | GTACGAAATGCGTGCCACAG | 402 |
| 3 | AGGGGATAGGTTAGTTTACTGTCA | AACCAGCTTCACAAAATGCAC | 275 |
| 4 | CACAGGAGGTATTTTAAGGCAGT | TGGGTGATAGTGAAGAAATCGGT | 332 |
| 5 | TGGTGTCAAAAACTTTAGCAGTTC | TCCAAAGAGAGCAGTAAGAAAAGT | 366 |
| 6 | CCTTTTCACCTCAAGTTTGTTG | GTGCAAGGTTGTGGTTGTG | 405 |
| 7 | AGGCTTTACTGTGCATAACTTACA | AATGTAAGTGAATGTCTGCCTGA | 329 |
| 8 | TGCACATTAGGCAAAAATTGGA | TCTAAACTCATTGTTGGGGCT | 311 |
| 9 | GGTGGTTGGTTATGTAGGAACA | CTACATAAAGTGCTCCCTGCC | 358 |
| 10 | TGTCCTTAAGTATGGTCATTGACTG | TGGACACAACAGAAAAAGCTAGA | 418 |
| 11 | AATAGCCAAGAGCACAGAGGC | ACATTGGCATACAAGTGTGGAGA | 265 |
| 12 | TCCGCTGTTTTAGGGTAGACA | GTTGTTTATCCAAGGTCACGTAGT | 609 |

* All sequences are provided in the 5' to 3' orientation. ORF = Open reading frame; bp = Base pairs.
Please note that exon 1 was too large to amplify in a single reaction and thus was subdivided into two smaller fractions (with sufficient overlap to analyze the entire sequence of exon 1, coding and non-coding).
Blue = newly designed primers.

## Appendix D: Expanded methodology for PCR reactions

### D.1 PCR recipe

| | Final Concentration | Volume per run |
|---|---|---|
| 2X GoTaq Green Master Mix* | 1X | 10 |
| 5 μM Forward and 5 μM Reverse primer (pre-mixed) | 0.25 μM | 1 |
| 100% DMSO** | 5% | 1 |
| distilled water | - | 3 |
| genomic DNA template (8 ng/μL) | 2 ng/μL | 5 |
| | | (Total volume: 20μL) |

* GoTaq Master Mix (Promega #M7123) includes: DNA polymerase, all dNTPs, $MgCl_2$ and other reaction buffers at optimal concentrations for DNA amplification by PCR.
**DMSO: dimethyl sulfoxide

### D.2 PCR conditions for *EZH2* amplification and sequencing

I. 94˚C, 2min
II. 94˚C, 45 sec
III. 55˚C, 1 min
IV. 72˚C, 1min (II-IV 36x) *
V. 72˚C, 10 min
VI. Keep at 4˚C

*Note that for DNA extracted from nail clippings, due to very low DNA concentration, number of cycles was increased to 42.

### D.3 PCR conditions for *NSD1* amplification and sequencing

For all primer pairs except exon 23.7:

I. 94˚C, 2min
II. 94˚C, 45 sec
III. 56˚C, 1 min
IV. 72˚C, 1min (II-IV 37x)
V. 72˚C, 10 min
VI. Keep at 4˚C

For exon 23.7:

I. 94˚C, 2min
II. 94˚C, 45 sec
III. 64˚C, 1 min

IV. 72˚C, 1min (II-IV 36x)
V. 72˚C, 10 min
VI. Keep at 4˚C

### D.4 PCR conditions for *EED* amplification and sequencing

For exon 1:

I. 94˚C, 2min
II. 94˚C, 45 sec
III. 63.2˚C, 30 sec
IV. 72˚C, 1min (II-IV 36x)
V. 72˚C, 10 min
VI. Keep at 4˚C

For exon 9:

I. 94˚C, 2min
II. 94˚C, 45 sec
III. 62.7˚C, 1 min
IV. 72˚C, 1min (II-IV 36x)
V. 72˚C, 10 min
VI. Keep at 4˚C

For all other exons (2-8 and 10-12): same protocol as for EZH2 sequencing.

## Appendix E: Sample pages from anonymized research sequencing report returned to referring physicians or families

Dr. William T. Gibson
Child and Family Research Institute
950 W. 28th avenue, A4-151, bay 17
VANCOUVER, BC, V5Z 4H4
Lab: 604-875-2000 ext.6783
wtgibson@cfri.ubc.ca

### Research Report: *EZH2* Sequence analysis
### Not for use in Clinical Management without Confirmation in a Clinical Testing Laboratory

| | |
|---|---|
| **Name:** Proband 5 | **Gender:** Male |
| **Study number:** - | **Date of birth:** - |
| **Sample type:** DNA (from blood) | **Received:** - |
| | **Reported:** - |
| **Indication for testing:** Physical features consistent with Weaver syndrome | |

**RESULT:** One mutation was found in the coding region of the *EZH2* gene.

| GENE | NUCLEOTIDE CHANGE | SEQUENCE POSITION | DNA ALTERATION | PROTEIN ALTERATION | ZYGOSITY | REFERENCE (dbSNP database) | INTERPRETATION |
|---|---|---|---|---|---|---|---|
| *EZH2* | chr7: 148,506,462C>T | exon 18 | c.2050C>T | p.Arg684Cys | Heterozygous | not reported | Likely pathogenic |

**INTERPRETATION:** The observed clinical features **may or may not** be caused by the observed variant in the *EZH2* coding sequence. Parents tested negative for this variant.

**OTHER FINDINGS**: Two non-coding variants were identified (see below). These variants are unlikely to have any phenotypic consequence.

| GENE | NUCLEOTIDE CHANGE | SEQUENCE POSITION | ZYGOSITY | REFERENCE (dbSNP database) | INTERPRETATION |
|---|---|---|---|---|---|
| *EZH2* | Del (A) chr7: 148,543,694 | Noncoding flanking region of exon 3 | Heterozygous | rs3214332 (variant frequency 24% in 1000 Genomes database) | Benign |
| *EZH2* | Del (C) chr7: 148,504,718 | Noncoding flanking region of exon 20 | Homozygous | rs3217095 (variant frequency 25.4% in 1000 Genomes database) | Benign |

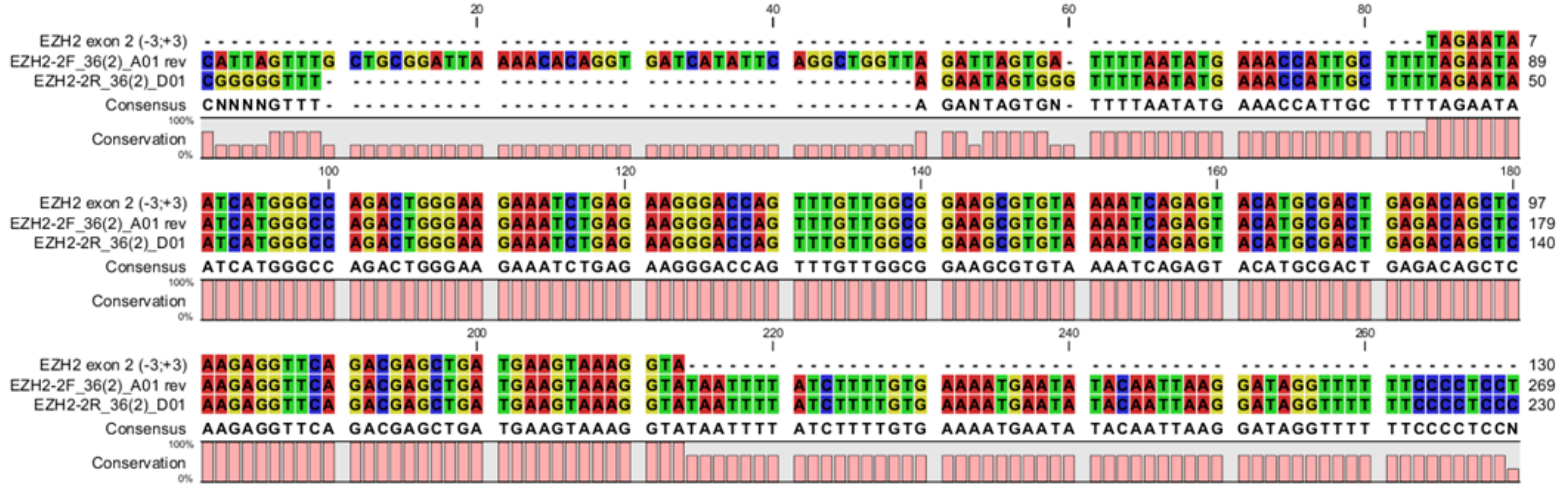
**METHODOLOGY:** All coding exons of *EZH2* (2-20) were PCR amplified and sequenced in both the forward and reverse directions from the individual's genomic DNA. Annotations are based on the Human Feb. 2009 (GRCh37/hg19) Assembly on the UCSC Genome Browser (http://genome.ucsc.edu/index.html), RefSeq summary NM_004456.4. Unknown variants were searched on a database of single nucleotide polymorphisms (SNPs) for previous reports (http://www.ncbi.nlm.nih.gov/snp).

**IMPORTANT NOTE:** This testing was performed in a RESEARCH setting. Though we make every effort to assure high-quality data, our protocols are considered "in development" and are not certified by Health Canada, the Canadian College of Medical Geneticists (CCMG), the College of American Pathologists (CAP), the International Laboratory Accreditation Cooperation (ILAC) or any similar body. Clinical-grade testing of the *EZH2* gene is now available to Canadian residents at the Molecular Genetics Laboratory at BC Children's Hospital (http://www.genebc.ca).

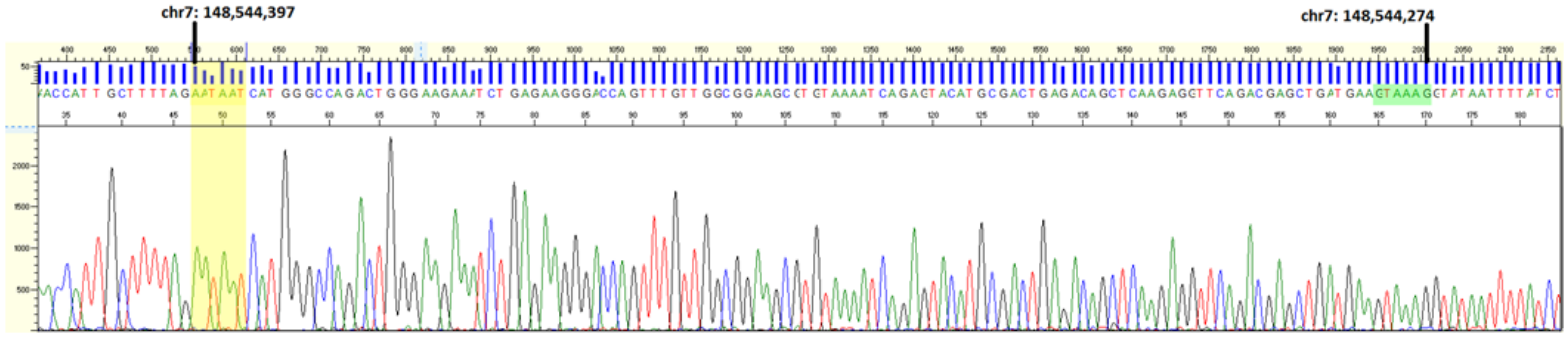
Detailed research report with sequence traces follows. Dr. Gibson is available for further discussion.

## EXON 2

Sequence: complete alignment of the coding region.

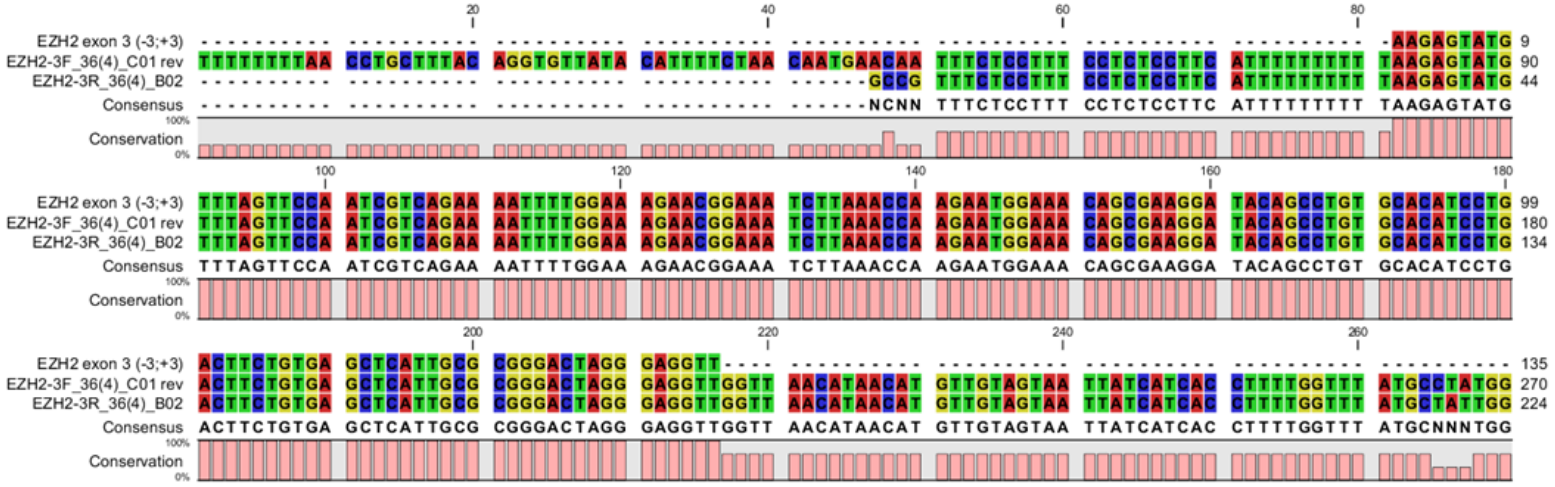



Sense strand: good quality peaks with minimal background, no mutations.

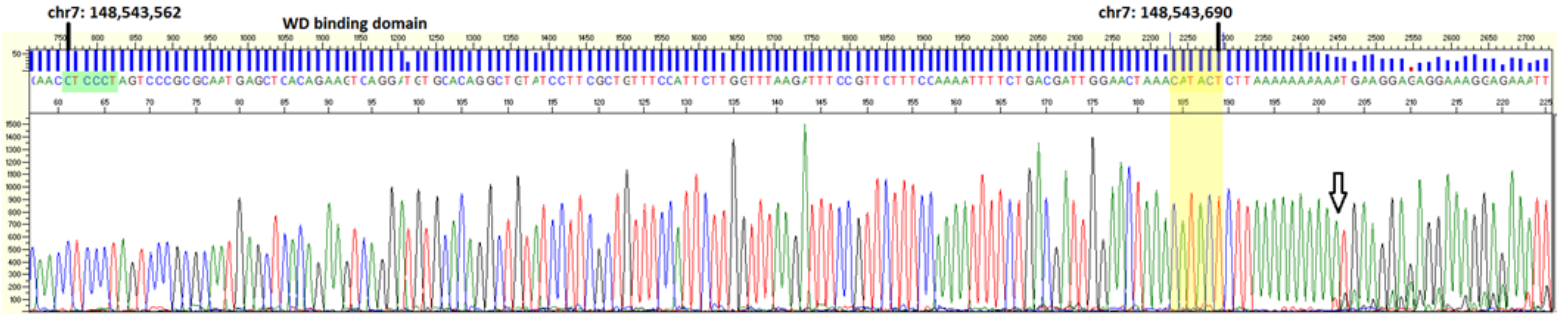


## EXON 3

Sequence: complete alignment of the coding region.

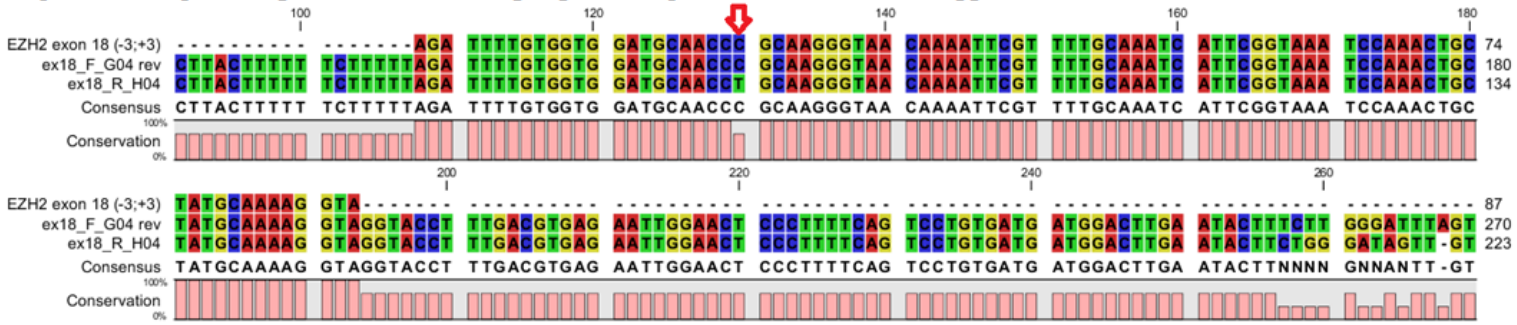



Antisense strand: good quality peaks with minimal background, no mutations in the coding region. Common intronic indel detected.

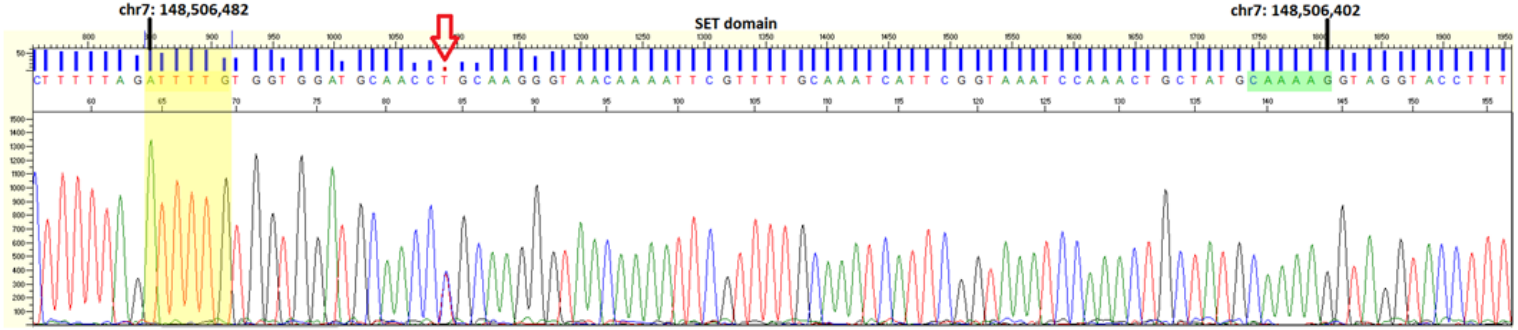


**EXON 18**

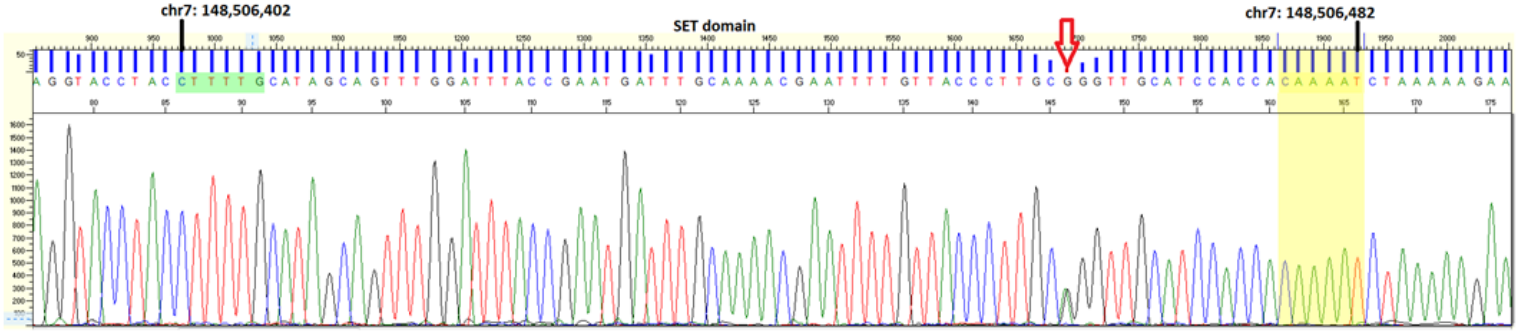
Sequence: complete alignment of the coding region except for one nucleotide, suggestive of variant/mutation.




Sense strand: good quality peaks, variant identified within the coding region (c.2050C>T; p.Arg684Cys).

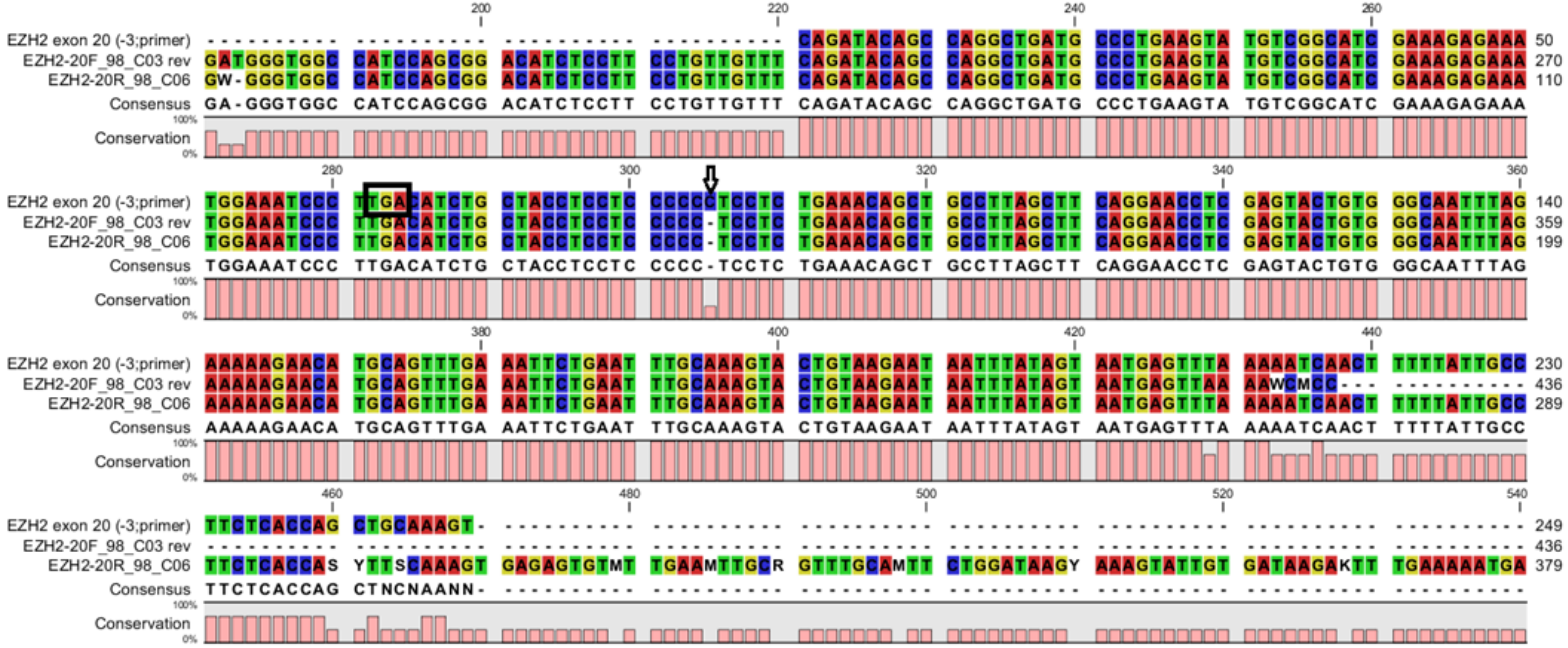



Antisense strand: variant confirmed. This variant has not been previously reported; pathogenicity will be inferred based on inheritance.

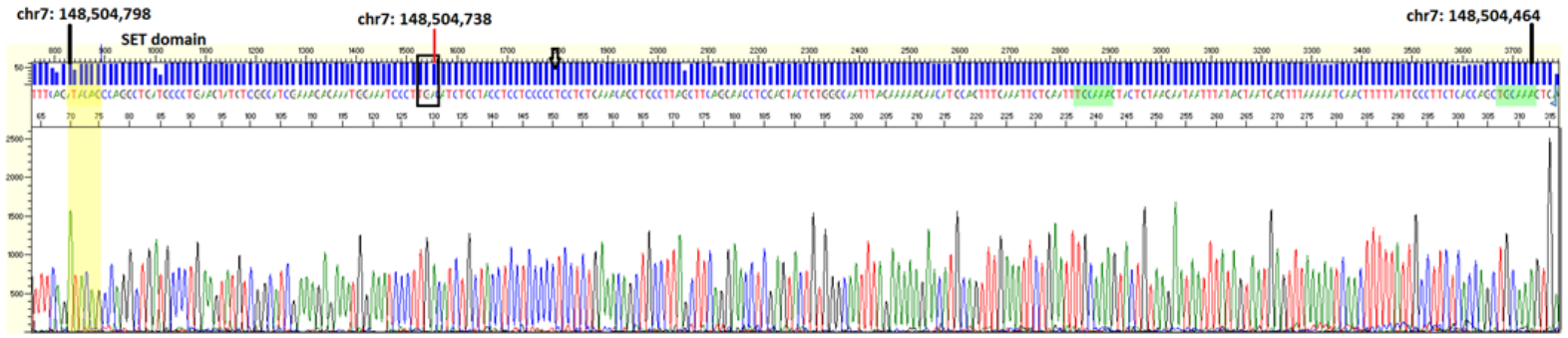


**EXON 20**

Sequence: complete alignment of coding region (stop codon noted with a black box). Intronic deletion detected in the 3' UTR.




Sense strand: good quality peaks with minimal background, no mutations in the coding region. Common intronic deletion confirmed.



## Appendix F: Health economics estimates - diagnostic workup for proband 5

| Disorder | Genetic/Genomic Test | Location Performing | Cost ($ USD) |
|---|---|---|---|
| Aneuploidies | Prenatal FISH for Chromosomes 13, 18, 21, X and Y, with Karyotype | BCCH Cytogenetics Lab | 540 |
| Deletion/Duplication Copy Number Variants | Affymetrix 6.0 Microarray | BCCH Cytogenetics Lab | 790 |
| Inversions, Translocations | Postnatal Karyotype | BCCH Cytogenetics Lab | 425 |
| Fragile X | *FMR1* PCR and Southern Blot | BCCH Molecular Genetics Lab | 300 |
| Simpson-Golabi-Behmel Syndrome | MLPA gene dosage testing for Glypican-3 and Glypican-4 and GPC3 sequencing | Hospital for Sick Children, Toronto, ON | 2020 |
| Oral-Facial-Digital Syndrome | *OFD1* sequencing and MLPA deletion-duplication analysis | Prevention Genetics, Marshfield, WI | 2310 |
| X-linked disorders | Maternal X-chromosome inactivation studies | Hospital for Sick Children, Toronto, ON | 400 |
| Sotos Syndrome | *NSD1* sequencing and MLPA deletion-duplication analysis | Prevention Genetics, Marshfield, WI | 2580 |
| Amino acidopathies | Plasma Amino Acids | BCCH Biochemical Diseases Lab | 80 |
| Peroxisomal Disorders including Zellweger syndrome | Very Long-Chain Fatty Acids | Kennedy Krieger Institute, Baltimore, MD | 90 |
| Carbohydrate-Deficient Glycoprotein Syndrome | Transferrin Isoelectric Focusing | BCCH Biochemical Diseases Lab | 90 |
| Weaver Syndrome | *EZH2* sequencing | BCCH Molecular Genetics Lab | 700 |
| **TOTAL** | | | **10,325** |

FISH = Fluorescence in situ Hybridization; MLPA = Multiplex Ligation-dependent Probe Amplification.

## Appendix G: Overlap of *EZH2* variants between Weaver syndrome and somatic cancers, according to the COSMIC database (July 2016)

| ***EZH2* variant (protein change)** | **Weaver syndrome proband(s)** | **COSMIC reference** | **Type of cancer** | **References** |
|---|---|---|---|---|
| p.Pro132Ser | 3, 4 | COSM133047 | Hematological: myelofibrosis | [305] |
| | | | Skin: malignant melanoma | nds |
| p.Pro132= | n/a | COSM5020984 | Soft tissue: hemangioblastoma | [451] |
| p.Tyr133Cys | 6,7 | nds | n/a | n/a |
| p.Tyr133His | n/a | COSM144172 | Hematological: acute lymphoblastic T-cell leukemia | [312] |
| p.Gly135Arg | 8 | COSM133044 | Hematological: myelofibrosis | [305] |
| p.Asp140Glu | 11 | nds | n/a | n/a |
| p.Tyr153del | 1 | nds | n/a | n/a |
| p.Tyr153Cys | n/a | COSM4384296 | Hematological: myelodysplastic syndrome | [452] |
| p.Ser669Asn | 9 | nds | n/a | n/a |
| p.Ser669Gly | n/a | COSM5427550 | Hematological: acute myeloid leukemia | nds |
| p.Ser669Arg | n/a | COSM4384253 | Hematological | [452] |
| p.Arg684Cys | 5 | COSM53005 | Hematological: chronic myelomonocytic leukemia and myelofibrosis | [264] |
| p.Arg684Ser | n/a | COSM600065 | Skin: malignant melanoma | [453] |
| p.Arg684His | n/a | COSM306072 | Hematological: myelodysplastic syndrome | [452] |
| | | | Hematological: acute lymphoblastic T-cell leukemia (x3) | [233,454] |
| | | | Large intestine: adenocarcinoma | nds |
| p.His694Tyr | 2 | COSM53040 | Hematological: chronic myelomonocytic leukemia | [304] |
| p.His694Arg | n/a | COSM87277, COSM4169591 | Hematological: mast cell neoplasm and myelodysplastic syndrome | [455,456] |
| p.Glu745Lys | 10 | COSM1087033 | Endometrium: carcinoma (x2) | nds |
| | | | Hematological: acute lymphoblastic T-cell leukemia | [454] |
| | | | Stomach: adenocarcinoma | nds |
| p.Glu745fs*>8 | n/a | COSM5487424 | Hematological: acute myeloid leukemia | nds |
| p.Glu745fs*>9 | n/a | COSM1319007 | Hematological: acute myeloid leukemia | nds |

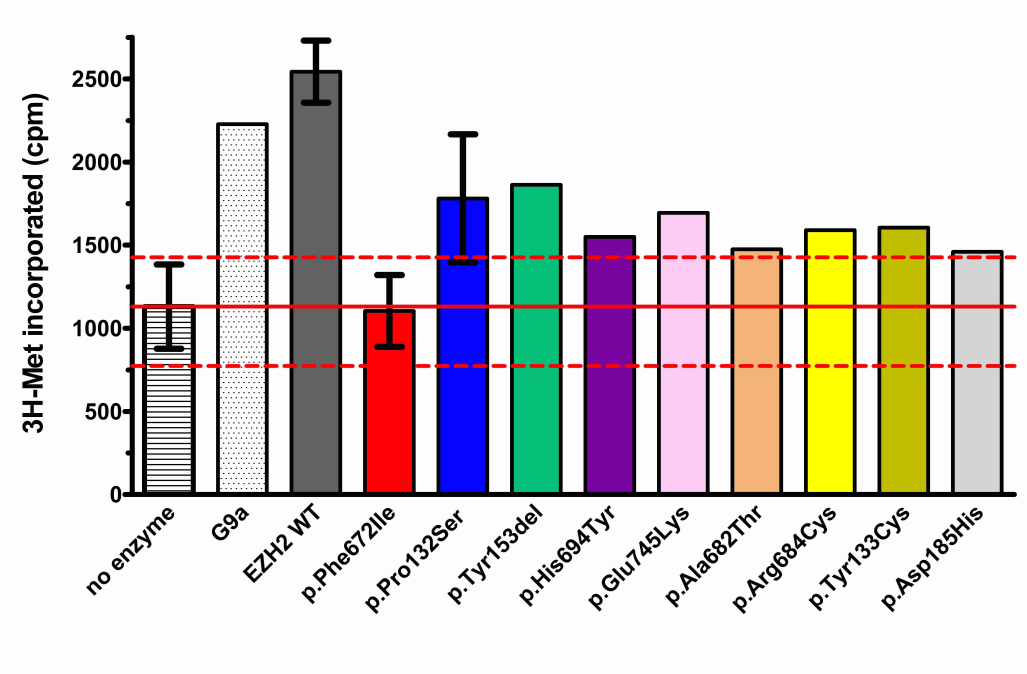
n/a = non applicable; nds = not described; black = variants described in our Weaver syndrome cohort; blue = variants in the same amino acids but described in somatic cancers only.

**Appendix H: Development of our *in vitro* histone peptide methyltransferase assay described in Chapter 3**

To begin, we repeated the standard reaction (described in section 3.2.2) using all complexes to be tested. G9a was once again used as a positive control enzyme. WT EZH2 was used as a positive control for PRC2 activity, and EZH2 p.Phe672Ile as a negative control, as described earlier. We incubated 250 ng of individual HMTase complexes separately with 0.67 μM $^{3}$H-SAM (Perkin Elmer) and 2 μg core histones, in 50 mM Tris-HCl, pH 9.0 and 0.5 mM DTT for 30 min at 30˚C in a 10 μl volume. Next, five or eight microlitres were spotted on a P81 square paper (Millipore), washed (three times with 10% tricholoracetic acid and once with 95% ethanol) to remove unincorporated $^{3}$H-Met, air-dried overnight, placed in a glass scintillation vial with 3 ml of scintillation fluid (ScintiSafe Econo1 SX20-5 or Scintisafe 30% SX23-5, Fisher Chemical) and counted on a 1900TR Liquid Scintillation Analyzer (Perkin Elmer) or LS6500 Multi-Purpose Scintillation Counter (Beckman Coulter). At this time, it was noted that counts were very low and often undistinguishable from background noise; as such, results are presented here as raw reads. Once again, our data suggest that the histone methyltransferase activity of mutant EZH2 is impaired *in vitro* (Appendix H.1).

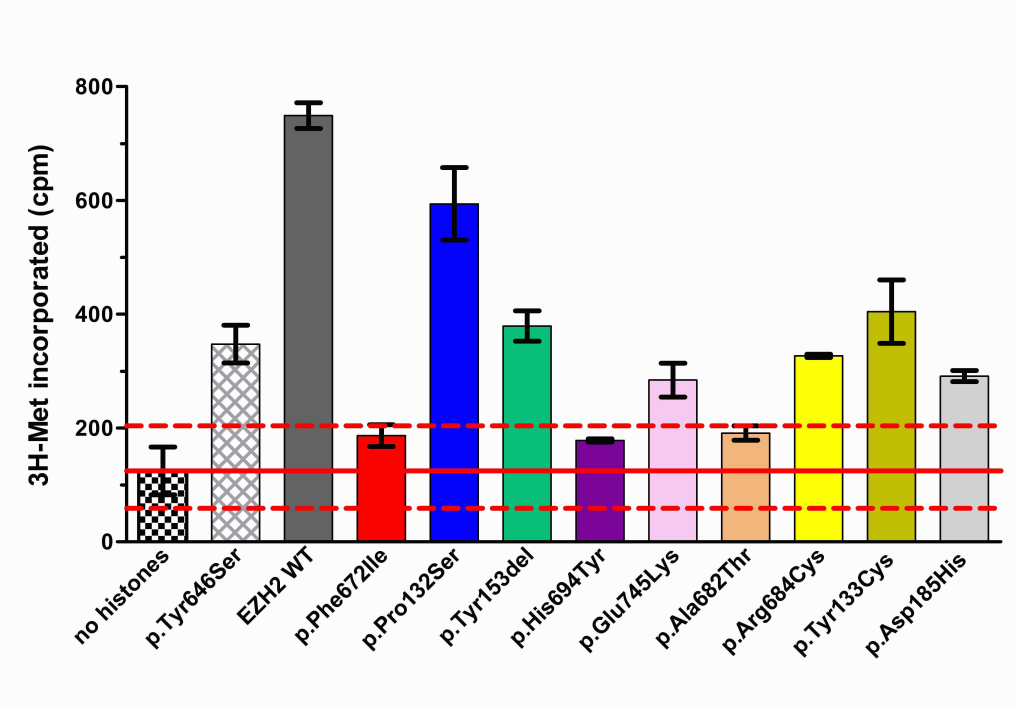
In addition, we carried out the same experiment with either an excess of core histones (Appendix H.2), an excess of $^{3}$H-SAM (Appendix H.3), or a longer reaction time (i.e. longer incubation at 30˚C) (Appendix H.4). All other reaction conditions remained the same within each experiment. As expected, based on the fact that this assay had been well established previously,[231,257,264] similar results were obtained under these varied conditions, still supporting a loss-of-function hypothesis for these mutant complexes.

## H.1 All EZH2 Weaver syndrome mutants appear to have impaired histone methyltransferase activity *in vitro*



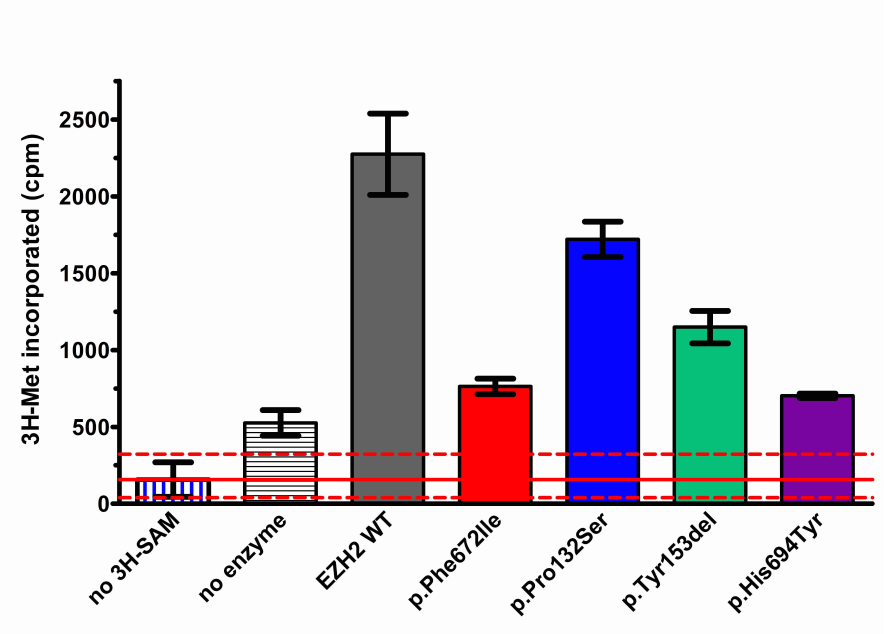

Histone methyltransferase reactions were performed using 2µg purified core histones and 0.67 µM $^{3}$H-S-Adenosyl-methionine ($^{3}$H-SAM). Each reaction was incubated with 250 ng of either wild-type (WT) or a mutant complex (or no enzyme controls). Active G9a (Millipore) was used as a further positive control. Histone methyltransferase activity was measured based on the incorporation of $^{3}$H-labeled methyl groups, represented in scintillation counts per minute (total counts). Error bars represent standard deviations of four independent replicates for WT and p.Phe672Ile, three independent replicates for the mutant p.Pro132Ser, and all 19 enzyme-free negative controls measured. There was only one measurement carried out for each of the remaining enzymatic complexes. The red lines represent mean background (solid line), and minimal or maximum background (dotted lines) observed in this experiment. Further statistical data on the background is presented in Appendix I.1.

## H.2 *In vitro* assay carried out with an excess of core histones still shows impaired histone methyltransferase activity



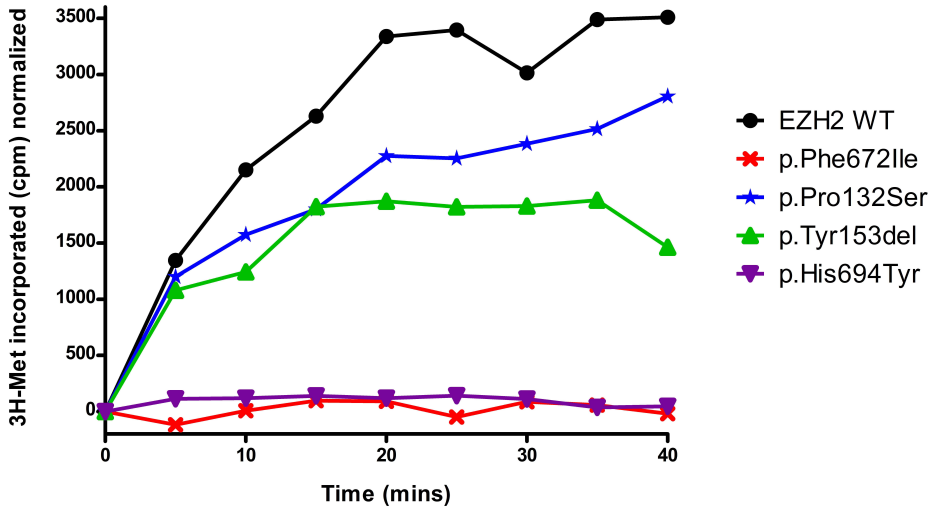

Histone methyltransferase reactions were performed with 250 ng of either wild-type (WT) or a mutant complex. Histone methyltransferase activity was measured based on the incorporation of $^{3}$H-labeled methyl groups, represented in scintillation counts per minute. Here, each reaction was incubated with **3 µg purified core histones** and 0.22 µM $^{3}$H-SAM, for 30 minutes at 30$^{0}$C. The p.Tyr646Ser mutant, commonly observed in cancer,[257] was added as an additional control. Values represent total counts. Error bars represent calculated standard deviations of two independent replicates for each complex. The red lines represent mean background (solid line), and minimal or maximum background (dotted lines) observed in this experiment. Incubation without histone substrates was also carried out to assess background radioactivity (including $^{3}$H-SAM bound non-specifically to the P81 square paper). Further statistical data on the background is presented in Appendix I.2.

## H.3 *In vitro* assay carried out with an excess of $^{3}$H-SAM still shows impaired histone methyltransferase activity




Histone methyltransferase reactions were performed with 250 ng of either wild-type (WT) or a mutant complex. Histone methyltransferase activity was measured based on the incorporation of $^{3}$H-labeled methyl groups, represented in scintillation counts per minute. Here, each reaction was incubated with 2 µg purified core histones and **1 µM $^{3}$H-SAM**, for 30 minutes at 30$^{0}$C. Values represent total counts. Error bars represent calculated standard deviations of two independent replicates for each mutant complex, and four replicates for WT. The red lines represent mean background (solid line), and minimal or maximum background (dotted lines) observed in this experiment. Incubation with no $^{3}$H-SAM was also carried out to assess background. Further statistical data on the background is presented in Appendix I.3.

## H.4 *In vitro* assay carried out with a longer reaction time still shows impaired histone methyltransferase activity




Histone methyltransferase reactions were performed with 250 ng of either wild-type (WT) or a mutant complex. Histone methyltransferase activity was measured based on the incorporation of $^{3}$H-labeled methyl groups, represented in scintillation counts per minute. Here, each reaction was incubated with 2 µg purified core histones and 0.67 µM $^{3}$H-SAM at 30$^{0}$C. Reactions were stopped at various time points (**0 to 40 minutes**) by taking an aliquot of the total reaction volume and spotting it onto P81 square paper. Values represented here were normalized by subtracting background counts (i.e. time zero) from the total counts. There was only one measurement for each enzymatic complex at each specific time point.

**Appendix I: Column statistics on the background reads measured with our *in vitro* histone peptide methyltransferase assay described in Chapter 3**

## I.1 Column statistics for graph in Appendix H.1

| | |
|---|---|
| Number of values | 19 |
| Sum | 21478 |
| Minimum | 773.0 |
| 25% Percentile | 882.0 |
| Median | 1249 |
| 75% Percentile | 1377 |
| Maximum | 1427 |
| 10% Percentile | 786.0 |
| 90% Percentile | 1418 |
| Mean | 1130 |
| Std. Deviation | 253.7 |
| Std. Error | 58.21 |
| Lower 95% CI of mean | 1008 |
| Upper 95% CI of mean | 1253 |
| Coefficient of variation | 22.45% |

## I.2 Column statistics for graph in Appendix H.2

| | |
|---|---|
| Number of values | 22 |
| Sum | 2744 |
| Minimum | 59.00 |
| 25% Percentile | 93.19 |
| Median | 111.5 |
| 75% Percentile | 169.3 |
| Maximum | 204.0 |
| 10% Percentile | 81.50 |
| 90% Percentile | 196.7 |
| Mean | 124.7 |
| Std. Deviation | 42.06 |
| Std. Error | 8.968 |
| Lower 95% CI of mean | 106.1 |
| Upper 95% CI of mean | 143.4 |
| Coefficient of variation | 33.73% |

## I.3 Column statistics for graph in Appendix H.3

| | |
|---|---|
| Number of values | 22 |
| Sum | 3457 |
| Minimum | 40.00 |
| 25% Percentile | 65.61 |
| Median | 99.00 |
| 75% Percentile | 293.0 |
| Maximum | 322.0 |
| 10% Percentile | 54.20 |
| 90% Percentile | 316.1 |
| Mean | 157.2 |
| Std. Deviation | 112.5 |
| Std. Error | 23.99 |
| Lower 95% CI of mean | 107.3 |
| Upper 95% CI of mean | 207.0 |
| Coefficient of variation | 71.60% |

## I.4 Column statistics for Figure 3-4

| | **me0 - me1** | **me1-me2** | **me2-me3** |
|---|---|---|---|
| Number of values | 24 | 30 | 54 |
| Sum | 35960 | 43960 | 79920 |
| Minimum | 1255 | 930.0 | 930.0 |
| 25% Percentile | 1393 | 1136 | 1299 |
| Median | 1510 | 1395 | 1461 |
| 75% Percentile | 1600 | 1655 | 1615 |
| Maximum | 1753 | 2406 | 2406 |
| 10% Percentile | 1279 | 1061 | 1104 |
| 90% Percentile | 1669 | 2274 | 1776 |
| Mean | 1498 | 1465 | 1480 |
| Std. Deviation | 138.8 | 399.3 | 309.6 |
| Std. Error | 28.32 | 72.89 | 42.13 |
| Lower 95% CI of mean | 1440 | 1316 | 1395 |
| Upper 95% CI of mean | 1557 | 1614 | 1565 |
| Coefficient of variation | 9.26% | 27.25% | 20.92% |

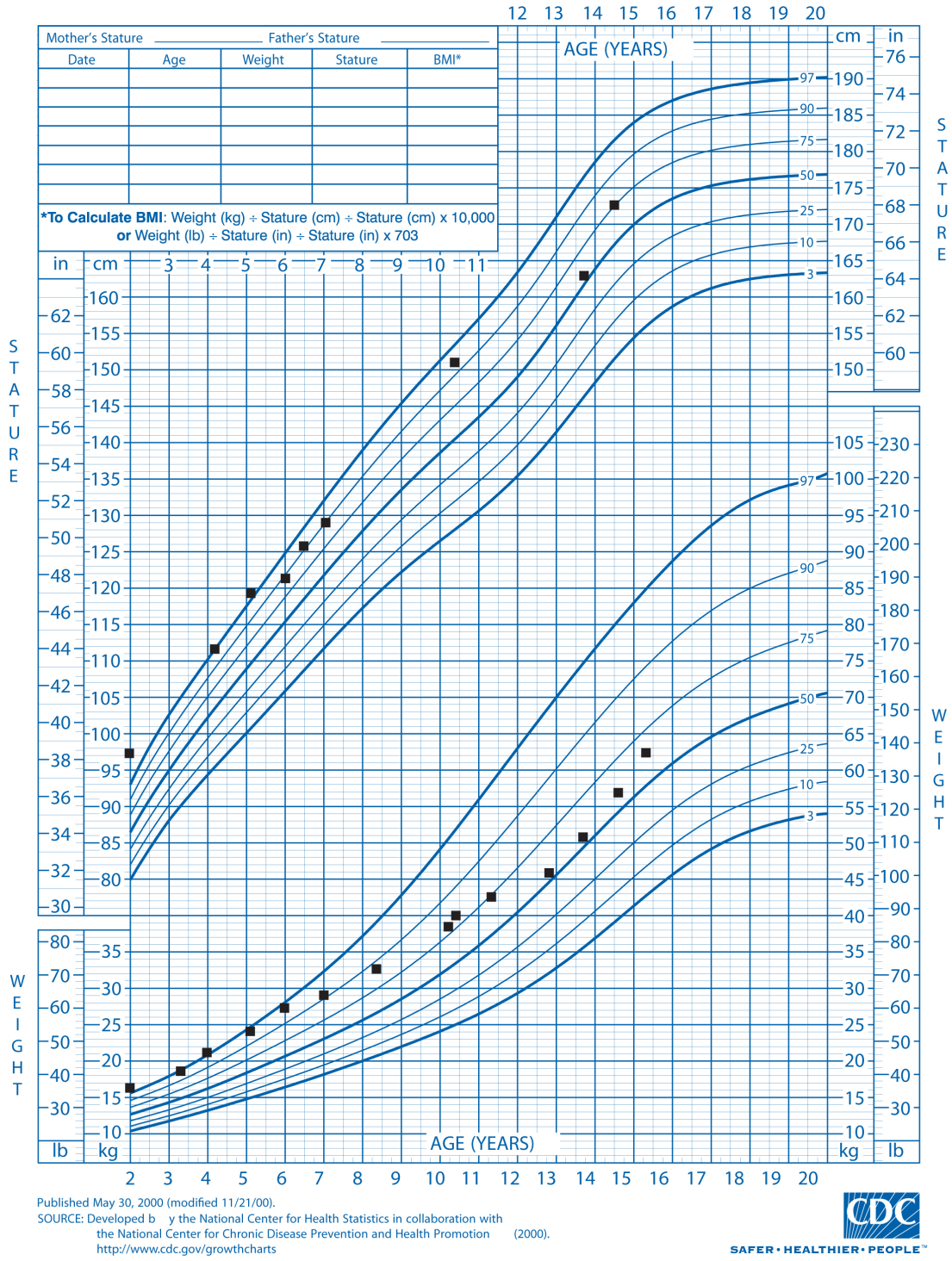
## Appendix J: Top candidate genes considered in the overgrowth gene discovery analysis, based on prior functional knowledge

| Genes (gene family) | Functional Reasoning | References |
|---|---|---|
| *EED* | essential member of PRC2 | [178,183,195] |
| *SUZ12* | essential member of PRC2 | [178,183,195] |
| *RBAP48/46* (RBBP) | alternative members of PRC2 | [178,183,195] |
| *AEBP2* | alternative member of PRC2 | [178,183] |
| *EZH1* | paralog of *EZH2*, alternative member of PRC2 | [183,195] |
| *NSD2/3* | related to *NSD1*, strong sequence similarity<br>(note: screening for mutations within these genes in a larger cohort did not yield promising results [457]) | [457] |
| *CBX2/4/6/7/8* (CBX) | alternative members of PRC1 | [183,195] |
| *PHC1/2/3* | alternative members of PRC1 | [183,195] |
| *RING1A/1B* | alternative members of PRC1 | [183,195] |
| *MEL18,NSPC1,BMI1* (PCGF) | alternative members of PRC1 | [183,195] |
| *SCML1/2/4* (SCML) | alternative members of PRC1 | [195] |
| *PHF1,MTF2,PHF19* (PCL) | bind PRC2 to recruit complex and stimulate H3K27 methylation | [183,195] |
| *UTX,JMJD3* (KDM) | H3K27 demethylases | [198] |
| *MLL1/2* (KMT) | lysine methyltransferases, SET-domain containing proteins | [458] |
| *SETD2/3/5/6/7/8* (SETD) | SET-domain containing proteins<br>(note: after the compilation of this list, *SETD2* mutations were shown to cause overgrowth by an independent group [161]) | [458] |
| *HDAC1/2* (HDAC) | histone deacetylases, directly bound by EED, components of the NuRD complex | [188,194,374] |
| *DNMT1/3A/3B* (DNMT) | DNA methyltransferases<br>(note: after the compilation of this list, *DNMT3A* mutations were shown to cause overgrowth by an independent group [160]) | [459] |
| *JARID2* | mediates PRC2 binding<br>(and later shown to be directly methylated by PRC2 [331]) | [196] |
| *AKT1* | human homolog of Akt, which has been shown to inhibit H3K27 by phosphorylating EZH2 | [208,209] |
| *CDK1/2* | phosphorylate EZH2 | [209] |
| *MAPK14* | also known as p38 alpha, phosphorylates EZH2 | [209] |
| *YY1* | plays a role in PRC2 recruitment | [460,461] |

## Appendix K: Summary of coverage check

| Gene | Related overgrowth syndrome | Overall coverage in exomes |
|---|---|---|
| *EZH2* | Weaver | Good coverage in all exomes |
| *NSD1* | Sotos | Good coverage in most exomes, intermediate coverage for a few exons in case 37; no concerns |
| *NFIX* | Sotos 2 | Good coverage in most exomes, intermediate coverage for a few exons in cases 50, 54, and 68; no concerns |
| *IGF2* | Beckwith-Wiedemann | Intermediate coverage in all exomes; no concerns |
| *CDKN1C* | Beckwith-Wiedemann | Intermediate coverage in all exomes; no concerns |
| *GPC3* | Simpson-Golabi-Behmel | Good coverage in all exomes |
| *FMR1* | Fragile X | Intermediate coverage in all exomes; no concerns |
| *FBN1* | Marfan | Good coverage in all exomes |
| *FBN2* | Beals | Good coverage in all exomes |
| *DNMT3A* | Tatton-Brown-Rahman | Good coverage in all exomes |
| *SETD2* | Luscan-Lumish | Good coverage in all exomes |
| *RNF125* | Tenorio | Good coverage in all exomes |
| *PDGFRB* | Kosaki overgrowth | Good coverage in all exomes |
| *PTEN* | various | Good coverage in all exomes |
| *NSD2/ WHSC1* | n/a (homolog of *NSD1*, good candidate) | Intermediate coverage in most exomes, good coverage for cases 17, 25 and 37; no concerns |
| *NSD3/ WHSC1L1* | n/a (homolog of *NSD1*, good candidate) | Good coverage in all exomes |
| *EED* | n/a (essential member of PRC2, good candidate) | Good coverage in all exomes |
| *SUZ12* | n/a (essential member of PRC2, good candidate) | Good coverage in all exomes |

**Appendix L: Height and weight measurements from 2 to 14 years of age for case 2 with *EED*-related overgrowth, plotted on a CDC clinical growth chart**



## Appendix M: Candidate variants identified and validated in the other completed exomes

| Case | Main phenotypic traits | Top candidate gene(s) | Predicted protein change | *In silico* functional prediction | Reasoning | Sanger validation results | Comments |
|---|---|---|---|---|---|---|---|
| 17 | overgrowth, macrocephaly, not very dysmorphic | ***TP53BP1*** (OMIM *605230) | Y1264* | n/a (STOP gained) | 53BP1 protein binds p53, which controls cell proliferation;[462] 53BP1 is also involved in double-strand break (DSB) repair by non-homologous end joining, which requires chromatin mobility: 53BP1 associates with chromatin surrounding DSBs and binds directly to nucleosomes by recognizing two histone modifications, H4K20me2 and H2AK15ub [463–465] | validated, but shown to be inherited from father who is deemed unaffected | phenotypic information was very limited at the time of prioritization for WES; recently obtained information remains unspecific |
| 25 | tall stature, macrocephaly, advanced bone age, round face with hypertelorism | ***SETD5*** (OMIM *615743) | P741R | PolyPhen: possibly damaging; SIFT: deleterious; PROVEAN: deleterious | chromatin regulator; and a recent paper [466] had described *SETD5 de novo* mutations associated with intellectual disability, mild dysmorphism and congenital abnormalities (OMIM #615761) | validated, but shown to be inherited from father who is deemed unaffected | *SETD5* variant unlikely to be causative and phenotype information is limited, which makes it difficult to prioritize other variants |
| 29 | tall stature, advanced bone age, severe ID, speech and developmental delay, autism, dysmorphism suggestive of Sotos, micrognathia, periventricular leucomalacia, cerebral palsy, perinatal hypoglycemia | ***SETD2*** (OMIM *612778) | T115A (or **T159A** in the longest isoform) | PolyPhen: benign; SIFT: tolerated; PROVEAN: neutral | chromatin regulator; and a recent paper [161] had described *SETD2* as an overgrowth gene causing Luscan-Lumish syndrome (OMIM #616831) | validated, but shown to be inherited from father | tall stature may be familial and father does not show any signs of generalized overgrowth; facial features do not overlap between proband and father thus this *SETD2* variant is unlikely to be causative – no other good candidates have been identified so far |
| 31 | tall stature, macrocephaly, Aspergers, ADHD, hypertelorism | *no variants validated so far* | n/a | n/a | n/a | n/a | phenotypic information was very limited at the time of prioritization; recently the patient has developed multiple |

| Case | Main phenotypic traits | Top candidate gene(s) | Predicted protein change | *In silico* functional prediction | Reasoning | Sanger validation results | Comments |
|---|---|---|---|---|---|---|---|
| | | | | | | | basal cell nevi, suggestive of Gorlin syndrome, but no variants appear to be consistent with this phenotype |
| 37 | overgrowth, macrocephaly, advanced bone age, ID, behavioural problems, long pointed face with dysmorphism, micrognathia, ventricular septal defect, toe clinodactyly, hernias | ***DNMT3A*** (OMIM *602769) | A382P (or **A571P** in the longest isoform) | PolyPhen: possibly damaging; SIFT: deleterious; PROVEAN: neutral | chromatin regulator; and a recent paper [160] had described *DNMT3A* as an overgrowth gene causing Tatton-Brown-Rahman syndrome (OMIM #615879), more recently supported by Tlemsani *et al*. (2016)[369] | validated and confirmed to be ***de novo*** | Tatton-Brown *et al*.[160] (and a subsequent study) state that only mutations within recognizable protein domains of *DNMT3A* cause overgrowth, and this variant is located within the ADD domain; however, the phenotypic information available for these cases is insufficient to determine whether it overlaps with our patient's phenotype – this variant remains our strongest candidate but further evidence is required |
| 50 | overgrowth, macrocephaly, mild ID, developmental and speech delay, autistic features, CT showing prominent Virchow spaces, bitemporal narrowing, ear tubes placed due to recurrent ear infections | *no variants validated so far* | n/a | n/a | n/a | n/a | all 3 siblings are tall and 1 has right foot 2/3 syndactyly and also had speech delay and needed ear tubes; phenotype is very unspecific |
| 54 | mild overgrowth in early years, borderline macrocephaly, delayed milestones, autistic and ADHD features, pointed chin, downslanting fissures, sleep related breathing disorder, inguinal hernia requiring surgery, large | *KDM6B* or ***JMJD3*** (OMIM *611577) | L501* | n/a (STOP gained) | chromatin regulator: H3K27 demethylase;[198] not previously known to be associated with overgrowth | not validated – false positive | phenotypic information was very limited at the time of prioritization; recently obtained information suggests this may not be a sporadic dominant disorder as many of the proband's features overlap with other family members |

| Case | Main phenotypic traits | Top candidate gene(s) | Predicted protein change | *In silico* functional prediction | Reasoning | Sanger validation results | Comments |
|---|---|---|---|---|---|---|---|
| | genitalia, hypothyroidism, 3-4 café au lait spots | | | | | | |
| 68 | overgrowth and respiratory distress at birth, but with microcephaly, round face with dysmorphic features, developmental delay, mild periventricular leukomalacia, patent ductus arteriosus and ventricular septal defect repaired by surgery, umbilical hernia, recurrent ear infections, cerebral palsy, feeding difficulties | ***JAG1*** (OMIM +601920) | R25Q | SIFT: tolerated; PROVEAN: neutral | mutations in *JAG1* are associated with Alagille syndrome (OMIM #118450); although our patient does not show all the features traditionally associated with this syndrome, we did observe some growth retardation, dysmorphism, and cardiac defects | validated, but shown to be inherited from mother who is deemed unaffected | phenotypic information was very limited at the time of prioritization; recently obtained information suggests this may not be an overgrowth disorder because the proband's healthy twin is now growing at a faster rate – no other candidates consistent with this phenotype have been identified |
| 70 | overgrowth, macrocephaly, advanced bone age, mild ID, food seeking behaviour, round face with dysmorphism but not Weaver-like, reactive lung disease, acanthosis nigricans, early puberty | *no variants validated so far* | n/a | n/a | n/a | n/a | phenotypic information was very limited at the time of prioritization; recently obtained information suggests this may not be a sporadic dominant disorder as sister shows similar features |
| 92 | mild overgrowth, macrocephaly, developmental and speech delay, ventricolomegaly, square face with pointed chin and frontal bossing, hyperterlorism, increased subcutaneous fat, | *FAM208A* or ***TASOR*** (OMIM *616493) | S294G | PolyPhen: benign; SIFT: deleterious; PROVEAN: neutral | chromatin regulator: component of HUSH silencing complex, maintains transcriptional silencing by recruiting SETDB1 for H3K9me3 deposition;[467] deletion of *Setdb1* in mouse ESCs reduced EZH2 binding as well as H3K27me3 levels at SETDB1 binding peaks [468] | validated, but shown to be inherited from mother who is deemed unaffected | this patient was one of the first recruited to the study but does not have classical Weaver dysmorphism – a new cause of disease might be expected; at this time, *CHD3* represents the most likely candidate |
| | | ***CHD3*** or Mi2-ALPHA | R985Q (or R1044Q in longest | PolyPhen: probably damaging; SIFT: | chromatin regulator: CHD3 is a core subunit of the NuRD complex, involved in transcriptional repression via | validated and confirmed to be ***de novo*** | |

| Case | Main phenotypic traits | Top candidate gene(s) | Predicted protein change | *In silico* functional prediction | Reasoning | Sanger validation results | Comments |
|---|---|---|---|---|---|---|---|
| | | (OMIM *602120) | isoform) | damaging; PROVEAN: deleterious | nucleosome remodeling and histone deacetylation;[374] somatic mutations in *CHD3* have been described in cancers[379,380] and constitutional mutations in *CHD4* have recently been shown to cause a syndromic presentation with overlapping features, including developmental delay, intellectual disability, macrocephaly and distinct facial dysmorphisms[388] | | |
| | | *AKAP8* or ***AKAP95*** (OMIM *604692) | P288R | SIFT: tolerated; PROVEAN: neutral | chromatin regulator: AKAP95 physically associates with MLL complexes (Trithorax) and promotes MLL's H3K4-specific methyltransferase activity;[469] mutations in MLL2 are associated with Kabuki syndrome,[470,471] a rare disorder characterized by growth retardation, microcephaly and typical dysmorphism (OMIM #147920) | validated, but shown to be inherited from mother who is deemed unaffected | |

n/a = non applicable; WES = whole exome sequencing; ID = intellectual disability; ADHD = attention deficit hyperactivity disorder; CT= computed tomography scan; ESCs = embryonic stem cells.